EVALUATION AND VISUAL REPRESENTATION OF ONLINE PRODUCT USING MACHINE LEARNING BASED SENTIMENTAL ANALYSIS

A Thesis

Submitted

In Partial Fulfillment of the Requirements

for the Degree of

Master of Technology

in

Advanced Computing and Data Science

Submitted by

SURABHI AGARWAL

(Roll No: 2001209004)

Under the Supervision of:

Mohd Usman Khan

(Assistant Professor)

Department of Computer Science & Engineering

Faculty of Engineering

INTEGRAL UNIVERSITY, LUCKNOW, INDIA

July, 2022

CERTIFICATE

This is to certify that **Mr. Surabhi Agarwal** (Roll No. 2001209004) has carried out the research work presented in the dissertation titled **"Evaluation and Visual Representation of Online Products Using Machine Learning based Sentiment Analysis"** submitted for partial fulfillment for the award of the **Master of Technology** in **Advanced Computing and Data Science Engineering** from **Integral University, Lucknow** under my supervision.

It is also certified that:

(i) This dissertation embodies the original work of the candidate and has not been earlier submitted elsewhere for the award of any degree/ diploma/certificate.

(ii) The candidate has worked under my supervision for the prescribed period.

(iii) The dissertation fulfills the requirements of the norms and standards prescribed by the University Grants Commission and Integral University, Lucknow, India.

(iv) No published work (figure, data, table etc) has been reproduced in the dissertation without express permission of the copyright owner(s).

Therefore, I deem this work fit and recommend for submission for the award of the aforesaid degree.

Mohd Usman Khan
(Assistant Professor)
Department of CSE,
Integral University, Lucknow
Date:
Place: Lucknow

DECLARATION

I hereby declare that the dissertation titled "**Evaluation and Visual Representation of Online Products Using Machine Learning based Sentiment Analysis**" is an authentic record of the research work carried out by me under the supervision of **Mohd Usman Khan**, Department of Computer Science & Engineering, for the period from August, 2021 to August, 2022 at Integral University, Lucknow. No part of this dissertation has been presented elsewhere for any other degree or diploma earlier.

I declare that I have faithfully acknowledged and referred to the works of other researchers wherever their published works have been cited in the dissertation. I further certify that I have not willfully taken other's work, para, text, data, results, tables, figures etc. reported in the journals, books, magazines, reports, dissertations, theses, etc., or available at web-sites without their permission, and have not included those in this M.Tech dissertation citing as my own work.

Date:

Signature _____

Name**: Surabhi Agarwal**

Enroll.No.: **2000101207**

RECOMMENDATION

On the basis of the declaration submitted by "**Surabhi Agarwal**", a student of M.Tech CSE (Advanced Computing and Data Science), successful completion of Pre presentation and the certificate issued by the supervisor, **Mohd Usman Khan**, Assistant Professor, Computer Science and Engineering Department, Integral University, the work entitled "**Evaluation and Visual Representation of Online Products Using Machine Learning based Sentiment Analysis**", submitted to department of CSE, in partial fulfillment of the requirement for award of the degree of Master of Technology Advanced Computing and Data Science, is recommended for examination.

_____	_____
Program Coordinator Signature	HOD Signature
Dr. Faiyaz Ahmad	**Mrs. Kavita Agrawal**
Dept. of Computer Science &Engineering	Dept. of Computer Science & Engineering
Date:_____	Date:_____

COPYRIGHT TRANSFER CERTIFICATE

Title of the Dissertation: **Evaluation and Visual Representation of Online Products Using Machine Learning based Sentiment Analysis**

Candidate's Name: **Surabhi Agarwal**

The undersigned hereby assigns to Integral University all rights under copyright that may exist in and for the above dissertation, authored by the undersigned and submitted to the University for the Award of the Master of Technology Advanced Computing and Data Science degree.

The Candidate may reproduce or authorize others to reproduce material extracted verbatim from the dissertation or derivative of the dissertation for personal and/or publication purpose(s) provided that the source and the University's copyright notices are indicated.

SURABHI AGARWAL

ACKNOWLEDGEMENT

I am highly grateful to the Head of Department of Computer Science and Engineering for giving me proper guidance and advice and facility for the successful completion of my dissertation.

It gives me a great pleasure to express my deep sense of gratitude and indebtedness to my guide **Mohd Usman Khan,** Assistant Professor, Department of Computer Science and Engineering, for his valuable support and encouraging mentality throughout the project. I am highly obliged to him for providing me this opportunity to carry out the ideas and work during my project period and helping me to gain the successful completion of my Project.

I am also highly obliged to the Head of department, **Mrs. Kavita Agrawal (**Head of Department, Department of Computer Science and Engineering) and P.G Program Coordinator **Dr. Faiyaz Ahmad,** Assistant Professor, Department of Computer Science and Engineering, for providing me all the facilities in all activities and for his support and valuable encouragement throughout my project.

My special thanks are going to all of the faculties for encouraging me constantly towork hard in this project.

I pay my respect and love to my parents and all other family members and friends for their help and encouragement throughout this course of project work.

SURABHI AGARWAL

TABLE OF CONTENT

CONTENT	PAGE NO.
Title Page	i
Certificate/s (Supervisor)	ii
Declaration	iii
Recommendation	iv
Copyright Transfer Certificate	v
Acknowledgement	vi
List of Tables	x
List of Figures	xi-xii
List of Abbreviations	xiii
Abstract	xiv
Chapter1: Introduction	1
1.1 Introduction	2
1.2 Sentiment Analysis	4
1.2.1 Why Sentiment Analysis ?	5
1.2.2 Methods of Sentiment Analysis	7
1.2.3 Types of Sentiment Analysis	9
1.2.4 Sentiment Analysis Scope	10
1.2.5 Sentiment Analysis Applications	10
1.2.6 Limitations of Sentiment Analysis	11
1.3 Machine Learning	12
1.3.1 What is Machine Learning Used For ?	13
1.3.2 Types of Machine Learning	14
1.3.3 Popular Machine Learning Algorithms	15
1.3.4 Limitations of Machine Learning	16
1.4 Word Cloud	18
1.4.1 What is Word Cloud Generator?	19
1.4.2 Why Use a Word Cloud Generator?	19
1.4.3 Where Word Clouds Excel?	19

1.4.4 Limitations of Word Cloud 20
1.5 Product Recommendation System 20
1.5.1 Benefits of Recommendation 21
1.5.2 Machine Learning Techniques 22
1.5.3 Limitations of Recommendation 23
1.6 Dissertation Outline 25
Chapter 2: Literature Review
2.1 Literature Survey 26
2.2 Comparative Literature Survey 30
2.3 Conclusion 32
Chapter 3: METHODOLOGY 33
3.1 Proposed Methodology 34
3.2 Problem Statement 35
3.3 Objective 36
3.4 Model Architecture 36
3.4.1 Data Collection 36
3.4.2 Data pre-processing 37
3.4.3 Data Classification 39
3.4.4 Data Exploration 41
3.4.5 Product Recommendation 44
3.5 Tools and APIs used 45
Chapter 4: PROPOSED WORK 49
4.1 Proposed Goal 50
4.2 Approach Flow Diagram 51
4.3 Extraction Of Reviews 52
4.4 Process of Filtering Reviews 53
4.5 Data Exploration 55
4.6 Classification of Reviews 57
4.7 Detection of Sentiments 66

Chapter 5: RESULT and ANALYSIS 67

5.1 Data Set 68

5.1.1 Scrapped Data 68

5.1.2 Cleaned Data 69

5.2 Word Cloud Generator 70

5.3 Accuracy Evaluation of ML Models 70

5.4 Polarity of Reviews 72

5.5 Product Recommendation 73

5.6 Analysis and Discussion on Result 73

Chapter 6: CONCLUSION and FUTURE WORK 74

6.1 Conclusion 75

6.2 Future Scope 76

References 80

Publications 85

Certificate of Publication 87

LIST OF TABLES

Table 2.1	Comparative Literature Survey	30
Table 3.1	Recommendation of Product	45
Table 5.1	Accuracy Evaluation	70

LIST OF FIGURES

Figure 1.1: Categories of Sentiments 5
Figure 1.2: Sentiment Analysis Techniques 8
Figure 1.3: Word Cloud 18
Figure 3.1: Sentiment Analysis Procedure Flowchart 36
Figure 3.2: Data Collection Methods 37
Figure 3.3: Data Preprocessing Procedure 38
Figure 3.4: Data Classification Method 40
Figure 3.5: Data Exploration Method 41
Figure 3.6: Formula for Naïve Bayesian Classifier 42
Figure 3.7: Graph for SVM Classifier 43
Figure 3.8: Graph for Logistic Regression 44
Figure 3.9: Imported Libraries 46
Figure 3.10: Google Sheet 47
Figure 4.1: Approach Flowchart 51
Figure 4.2: Scrapped Data 52
Figure 4.3: Imported Data 53
Figure 4.4: Data after Preprocessing 54
Figure 4.5: Word Cloud for Review Title 55
Figure 4.6:Word Cloud for Review Content 55
Figure 4.7: Balance of Dataset 56
Figure 4.8: Accuracy of Machine Learning Models 58
Figure 4.9: Data for Polarity of Review 59
Figure 4.10: Polarity of Review Title (Histogram) 60
Figure 4.11: Polarity of Review Content (Histogram) 61
Figure 4.12: Polarity of Review Title (Box Plot) 62
Figure 4.13: Polarity of Review Content (Box Plot) 63
Figure 4.14: Distribution of Title Subjectivity Score 64
Figure 4.15: Distribution of Content Subjectivity Score 65
Figure 4.16: Data after VADER Analysis 66
Figure 5.1: Raw Data 68
Figure 5.2: Cleaned Data 69
Figure 5.3: Word Cloud 70
Figure 5.4: Accuracy of Machine Learning Models 71

Figure 5.5: Polarity (Pie Chart) 72
Figure 5.6: Polarity (Histogram) 73

LIST OF ABBREVIATIONS AND SYMBOLS

SVM	Support Vector Machine
NB	Naïve Bayes
NLP	Natural Language Processing
E-Commerce	Electronic Commerce
e.g	Example
ML	Machine Learning
VoC	Voice of Customer
LR	Linear Regression
KNN	K-Nearest Neighbour
AI	Artificial Intelligence
LSTM	Long Short-term Memory
AUC	Area Under Curve
RNN	Recurrent Neural Network
BoW	Bag of Words
URL	Uniform Resource Locator

ABSTRACT

Owing to the rise in demand for e-commerce with people preferring online buying of goods and products, there is huge amount information being shared. The e-commerce websites are carrying very big volume of data. Also, social media helps a great hand in sharing of this information. This has greatly influenced consumer preferences all over the world. Due to the intense reviews provided by the customers, there is a feedback environment being developed for helping customers buy the right product and guiding companies to enhance the features of product suiting consumer's demand. The only disadvantage of availability of this large volume of data is its range and its structural non-uniformity. The customer finds it difficult to precisely find the review for a particular feature of a product that user intends to buy. Also, there is a mixture of positive and negative reviews thereby making it difficult for customer to find a satisfactory response. Also these reviews suffer from fake reviews from fake users. So to avoid this confusion and make this review system more transparent and user friendly we propose a technique to extract feature based opinion from a diverse hub of reviews and processing it further to differentiate it with respect to the aspects of the product and further categorize it into positive and negative reviews using machine learning based approach. Decision making on both individual and organizational level is always accompanied by the search of other's opinions on the same because data holds expressed opinions and sentiments. The volume, variety and velocity are the key properties of this data. There are several tools and algorithms available to perform sentiment detection and analysis, which are better than unconventional, time consuming and error prone methods used earlier.

CHAPTER-1

INTRODUCTION

1.1 INTRODUCTION

Presently, very huge amount of data is available on internet. This data holds expressed opinions and sentiments. The volume, variety and velocity are the key properties of this data. Decision making on both individual and organizational level is always accompanied by the search of other's opinions on the same. With the tremendous establishment of opinion rich resources like product reviews, feedbacks are proved to be the most essential and valuable resources to market. Sentiment Analysis is an application of Natural Language Processing (NLP), also known as emotion extraction or opinion mining or text mining. It helps to understand the human decision making, categorizing, analyzing and extracting meaningful information in order to understand opinions of consumers. There are several tools and algorithms available to perform sentiment detection and analysis, which are better than unconventional, time consuming and error prone methods used earlier.

The advancement of electronic commerce with growth in internet and network technologies has led customers to move to online retail platforms such as Amazon, Walmart, etc. People often rely on customer reviews of products before they buy online. These reviews are often rich in information describing the products and their quality. Customers choose to compare between various products and brands based on whether an item has a positive or negative review. These reviews act as a feedback mechanism for the seller. Through this medium, sellers strategize their future sales and the areas where the product or services needs improvement.

The intense competition to attract and maintain customers online is compelling businesses to implement novel strategies to enhance the customer experiences. It is becoming necessary for companies to examine customer reviews on online platforms such as Amazon to understand better how customers rate their products and services. The purpose of this study is to investigate how companies can conduct sentiment analysis based on Amazon reviews to gain more insights into

customer experiences. The dataset selected for this research consists of customer reviews of Amazon products, which enables a business person to gain insights on customer reviews regarding specific product and services. The study will enable companies to pinpoint the reasons for positive and negative reviews, followed by implementing effective strategies to address them accordingly. The aim of this research is to help companies to use sentiment analysis to understand customer experiences and customers to understand whether a particular product is to be purchased or not.

In the recent years E-Commerce has exploded everywhere in the world, and majority of the population is preferring to buy products through these websites. Consequently large amount of data in the form of reviews is produced which helps prospective buyers to choose the right product. Furthermore these reviews contain opinionated contents which can be useful for the company to identify the areas which need to be enhanced. However it is impractical for the user to read each and every review about the product. Moreover, reading only few reviews may present a biased idea about the product. It is quite possible that some of the reviews lack credible sources, which the users have no means to differentiate. Besides the reviews and ratings provided do little to assess the specific features of the product. Due to all the above constraints, the user is unable to make a fully informed decision about the product. Opinion mining also known as sentiment analysis can be used to extract customer reviews from different sources on the internet. This technique implements various algorithms to analyze the corpus of data and make sense out of it. This technique helps to identify the orientation of a sentence thereby recognising the element of positivity or negativity in it. Automated opinion mining can be implemented through a machine learning based approach. Opinion mining uses natural language processing to extract the subjective information from the data (in this case it’s customer reviews).

The enormous amount of competition to attract and maintain customers online is fascinating businesses to implement novel strategies to enhance the customer experiences. It is becoming compulsory for companies to examine customer reviews on online platforms such as Amazon to understand better how customers rate their products and services. The purpose of this study is to investigate how companies can conduct sentiment analysis based on Amazon reviews to gain more intuitions into customer experiences. The dataset selected for this research consists of customer reviews of Amazon products, which enables a business person to gain insights on customer reviews regarding specific product and services. The study will enable companies to pinpoint the reasons for positive and negative reviews, followed by implementing effective strategies to address them accordingly. The aim of this research is to help companies to use sentiment analysis to understand customer experiences and customers to understand whether a particular product is to be purchased or not.

1.2 SENTIMENT ANALYSIS

Sentiment Analysis or opinion mining is one of the important tasks of NLP (Natural Language Processing) that has acquired much attention in recent years. The sentiment is a feeling, thought expression, or judgment and using sentiment analysis one can study the target audience's sentiments towards a particular product. It's a form of text analysis that senses polarity (e.g. a positive or negative opinion) within whole text, sentence, paragraph or phrase. Knowing people's emotions is important for companies and first time buyers because consumers can communicate their thoughts and feelings more freely. With the technological improvements in the field of machine learning and automation, companies can create systems th at automatically analyzes customer's feedback, survey responses and social media interactions. In this way, companies can listen to their customers closely and customize goods and services to suit the needs of their customers.

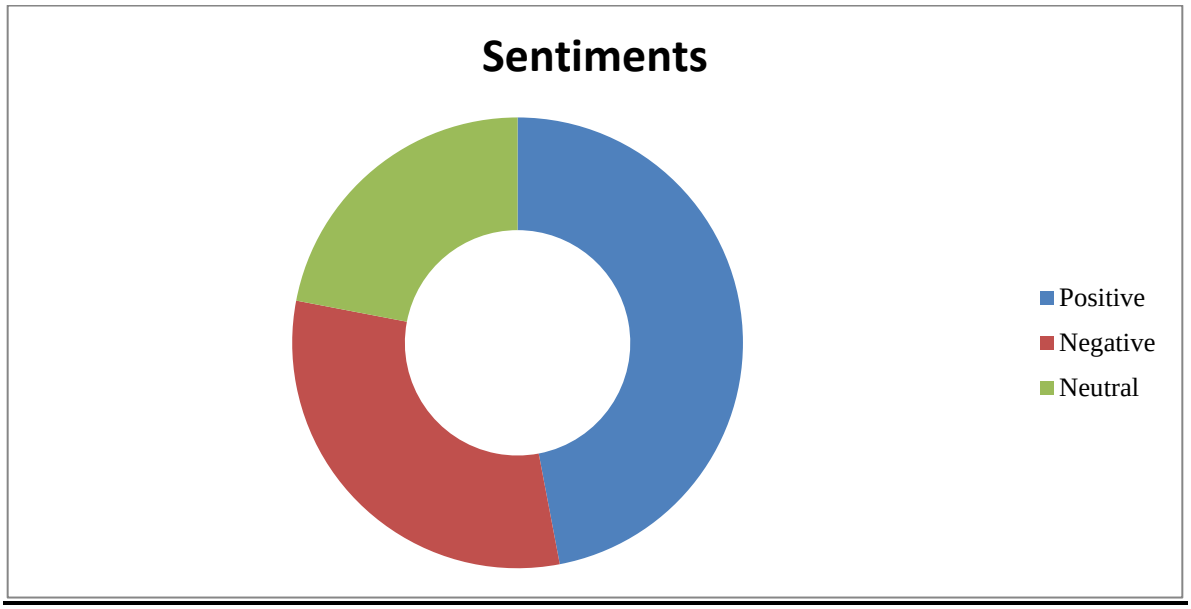


Fig. 1.1 Categories of Sentiments

1.2.1 Why Sentiment Analysis?

Simply reading a post will let you identify whether the author had a positive stance or a negative stance on the topic – but that's if you're well versed in the language. However, a computer has no concept of naturally spoken language – so, we need to break down this problem into mathematics (the language of a computer). It cannot simply deduce whether something contains joy, frustration, anger, or otherwise – without any context of what those words mean. Sentiment Analysis solves this problem by using Natural Language Processing. Basically, it recognizes the necessary keywords and phrases within a document, which eventually help the algorithm to classify the emotional state of the document.

Data Scientists and programmers write an application which feeds the documents into the algorithm and stores the results in a way which is useful for clients to use and understand. Keyword spotting is one of the simplest techniques and leveraged widely by Sentiment Analysis algorithms. The fed Input document is thoroughly scanned for the obvious positive and negative words like "sad", "happy", "disappoint", "great", "satisfied", and such.

There are a number of Sentiment Analysis algorithms, and each has different libraries of words and phrases which they score as positive, negative, and neutral. These libraries are often called the “bag of words” by many algorithms.

Although this technique looks perfect on the surface, it has some definite shortcomings. Consider the text, “The service was horrible, but the ambiance was awesome!” Now, this sentiment is more complex than a basic algorithm can take into account – it contains both positive and negative emotions. For such cases, more advanced algorithms were devised which break the sentence on encountering the word “but” (or any contrastive conjunction). So, the result becomes “The service was horrible” and “But the ambiance was awesome.”

This sentence will now generate two or more scores (depending on the number of emotions present in the statement). These individual scores are consolidated to find out the overall score of a piece. In practice, this technique is known as Binary Sentiment Analysis.

No Machine Learning algorithm can achieve a perfect accuracy of 100%, and this is no different. Due to the complexity of our natural language, most of the sentiment analysis algorithms are only 80% accurate, at best.

With everything shifting online, Brands have started giving utmost importance to Sentiment Analysis. Honestly, it’s their only gateway to thoroughly understanding their customer-base, including their expectations from the brand. Social Media listening can help organisations from any domain understand the grievances and concerns of their customers – which eventually helps the organisations scale up their services. Sentiment Analysis helps brands tackle the exact problems or concerns of their customers.

According to some researchers, Sentiment Analysis of Twitter data can help in the prediction of stock market movements. Researches show that news articles and social media can hugely influence the stock market. News with overall positive sentiment has been observed to relate to a

large increase in price albeit for a short period of time. On the other hand, negative news is seen to be linked to a decrease in price but with more prolonged effects. Ideally, sentiment analysis can be put to use by any brand looking to:

- Target specific individuals to improve their services.
- Track customer sentiment and emotions over time.
- Determine which customer segment feels more strongly about your brand.
- Track the changes in user behavior corresponding to the changes in your product.
- Find out your key promoters and detractors.

Clearly, sentiment analysis gives an organisation the much-needed insights on their customers. Organizations can now adjust their marketing strategies depending on how the customers are responding to it. Sentiment Analysis also helps organisations measure the ROI of their marketing campaigns and improve their customer service. Since sentiment analysis gives the organisations a sneak peek into their customer's emotions, they can be aware of any crisis that's to come well in time – and manage it accordingly.

1.2.2 Methods of Sentiment Analysis

Machine Learning based -

You're aware of the basic workings of any Machine Learning algorithms. The same route by followed in ML-based sentiment analysis algorithms as well. These algorithms require you to create a model by training the classifier with a set of example. This ideally means that you must gather a dataset with relevant examples for positive, neutral, and negative classes, extract these features from the examples and then train your algorithm based on these examples. These algorithms are essentially used for computing the polarity of a document.

Lexicon based -

As the name suggests, these techniques use dictionaries of words. Each word is annotated with its emotional polarity and sentiment strength. This dictionary is then matched with the document to calculate its overall polarity score of the document. These techniques usually give high precision but low recall. There is no "best" choice out of the two, your choice of method should depend solely on the problem at hand. Lexical algorithms can achieve near-perfect results, but, they require using a lexicon – something that's not always available in all the languages. On the other hand, ML-based algorithms also deliver good results, but, they require extensive training on labeled data.

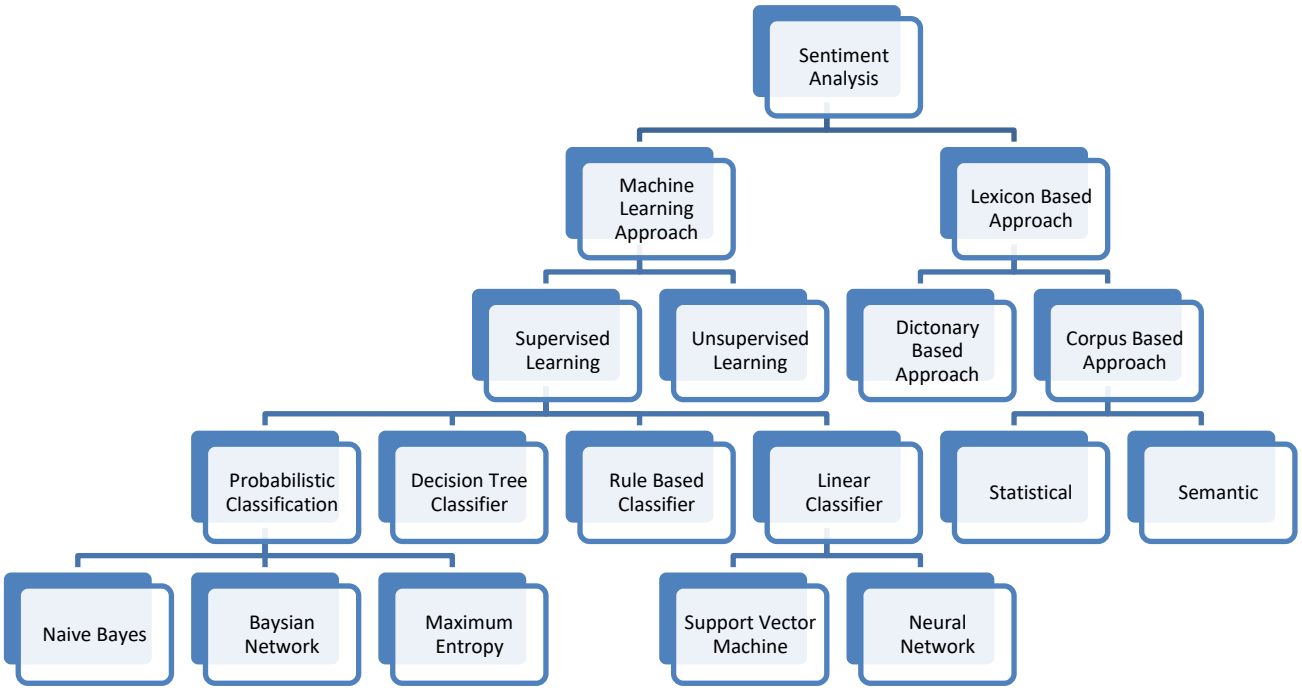


Fig. 1.2 Sentiment Analysis Techniques

1.2.3 Types of Sentiment Analysis

Fine-Grained -

This analysis gives you an understanding of the feedback you get from customers. You can get precise results in terms of the polarity of the input. However, the process to understand this can be more labor and cost-intensive as compared to other types.

Emotion Detection -

This is a more sophisticated way of identifying the emotion in a piece of text. Lexicons and machine learning are used to determine the sentiment. Lexicons are lists of words that are either positive or negative. This makes it easier to segregate the terms according to their sentiment. The advantage of using this is that a company can also understand why a customer feels a particular way. This is more algorithm-based and might be complex to understand at first.

Aspect based -

This type of sentiment analysis is usually for one aspect of a service or product. For example, if a company that sells televisions uses this type of sentiment analysis, it could be for one aspect of televisions – like brightness, sound, etc. So they can understand how customers feel about specific attributes of the product.

Intent based -

This is a deeper understanding of the intention of the customer. For example, a company can predict if a customer intends to use the product or not. This means that the intention of a particular customer can be tracked, forming a pattern, and then used for marketing and advertising.

Different methods are used for these different types of sentiment analysis – while one is rule-based, the other is automatic. Rule-based sentiment analysis is more rigid and might not always be accurate. It involves the natural language processing (NLP) routine. On the other hand,

automatic sentiment analysis is more detailed and in-depth. Machine learning is used to decode the feedback provided by each customer. So, there is more precision and flexibility here.

1.2.4 Sentiment Analysis Scope

- Document level sentiment analysis gets the sentiment of a total report or passage.
- Sentence level sentiment analysis gets the sentiment of a single sentence.
- Sub-sentence level sentiment analysis acquires the sentiment of sub- expressions inside a sentence.

1.2.5 Sentiment Analysis Applications

Customers contact businesses through multiple channels, and it can be hard for teams to stay on top of all this incoming data. With sentiment analysis tools, however, you can automatically sort your data as and when it filters into your help desk. Let's take a look at the most popular applications of sentiment analysis:

- Social media monitoring
- Customer support
- Brand monitoring and reputation management
- Listen to voice of the customer (VoC)
- Listen to your employees
- Product analysis
- Market research
- Future Sales Prediction
- Competitor research

1.2.6 Limitations of Sentiment Analysis

More or less every major brand these days relies heavily on social media listening to improve the overall customer experience. If you're one of the interested souls and want to explore this topic in further depth, we recommend you go through the various kinds of algorithms (the ones we displayed in a graphic earlier) and implementations of Sentiment Analysis in more detail.

Sentiment analysis is a great way to understand what the general opinion of the public is, specific to a company or a product. However, it has its own set of challenges and limitations, which can be overcome if it is used efficiently. Sometimes, it is difficult to understand the tone of the feedback, especially if there are irony and sarcasm involved.

Moreover, some algorithms are complicated and may not produce very insightful results. However, sentiment analysis is an excellent way to get unbiased opinions from customers about several things. It can help companies in a lot of aspects, especially when it comes to marketing and advertising or market research.

Sentiment analysis tools can identify and analyse many pieces of text automatically and quickly. But computer programs have problems recognizing things like sarcasm and irony, negations, jokes, and exaggerations - the sorts of things a person would have little trouble identifying. And failing to recognize these can skew the results.

'Disappointed' may be classified as a negative word for the purposes of sentiment analysis, but within the phrase "I wasn't disappointed", it should be classified as positive.

We would find it easy to recognize as sarcasm the statement "I'm really loving the enormous pool at my hotel!", if this statement is accompanied by a photo of a tiny swimming pool; whereas with short sentences and pieces of text, for example like those you find on Twitter especially, and automated sentiment analysis tool probably would not, and would most likely classify it as an example of positive sentiment.

Sometimes on Facebook, there might not be enough contexts for a reliable sentiment analysis. However, in general, Twitter has a reputation for being a good source of information for sentiment analysis, and with the new increased word count for tweets it's likely it will become even more useful.

So, automated sentiment analysis tools do a really great job of analysing text for opinion and attitude, but they're not perfect. When you're using a tool like Typely to analyse your text to see if it conveys the sentiment you want for your readers/audience, combine the results it gives you with your human judgments to identify anything the tool may not be able to easily determine. Typically highlights phrases in your text by positive and negative sentiment, making it super easy for you to see where your document is either expressing exactly the sentiments you want it to, or where you may need to make some changes.

1.3 MACHINE LEARNING

Machine Learning, as the name says, is all about machines learning automatically without being explicitly programmed or learning without any direct human intervention. This machine learning process starts with feeding them good quality data and then training the machines by building various machine learning models using the data and different algorithms. The choice of algorithms depends on what type of data we have and what kind of task we are trying to automate. Machine Learning is a subset or specific application of Artificial intelligence that aims to create machines that can learn autonomously from data. Machine Learning is specific, not general, which means it allows a machine to make predictions or take some decisions on a specific problem using data. Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics.

As big data continues to expand and grow, the market demand for data scientists will increase, requiring them to assist in the identification of the most relevant business questions and subsequently the data to answer them.

1.3.1 What is Machine Learning Used For?

Machine Learning is used in almost all modern technologies and this is only going to increase in the future. In fact, there are applications of Machine Learning in various fields ranging from smartphone technology to healthcare to social media, and so on.

Smartphones use **personal voice assistants** like Siri, Alexa, Cortana, etc. These personal assistants are an example of ML-based speech recognition that uses Natural Language Processing to interact with the users and formulate a response accordingly. Machine Learning is also used in social media. Let's take Facebook's **'People you may know'** as an example. It is mind-boggling how social media platforms can guess the people you might be familiar with in real life. And they are right most of the time!!! This is done by using Machine Learning algorithms that analyze your profile, your interests, your current friends, and also their friends and various other factors to calculate the people you might potentially know.

Machine Learning is also very important in **healthcare diagnosis** as it can be used to diagnose a variety of problems in the medical field. For example, Machine Learning is used in oncology to train algorithms that can identify cancerous tissue at the microscopic level at the same accuracy as trained physicians. Another famous application of Machine Learning is **Google Maps**. The Google Maps algorithm automatically picks the best route from one point to another by relying on the projections of different timeframes and keeping in mind various factors like traffic jams, roadblocks, etc. In this way, you can see that the applications of Machine Learning are limitless. If anything, they are only increasing and Machine Learning may one day be used in almost all fields of study!

1.3.2 Types of Machine Learning

Supervised Machine Learning

Imagine a teacher supervising a class. The teacher already knows the correct answers but the learning process doesn't stop until the students learn the answers as well. This is the essence of Supervised Machine Learning Algorithms. Here, the algorithm learns from a training dataset and makes predictions that are compared with the actual output values. If the predictions are not correct, then the algorithm is modified until it is satisfactory. This learning process continues until the algorithm achieves the required level of performance. Then it can provide the desired output values for any new inputs.

Unsupervised Machine Learning

In this case, there is no teacher for the class and the students are left to learn for themselves! So for Unsupervised Machine Learning Algorithms, there is no specific answer to be learned and there is no teacher. In this way, the algorithm doesn't figure out any output for input but it explores the data. The algorithm is left unsupervised to find the underlying structure in the data in order to learn more and more about the data itself.

Semi-Supervised Machine Learning

The students learn both from their teacher and by themselves in Semi-Supervised Machine Learning and you can guess that from the name itself! This is a combination of Supervised and Unsupervised Machine Learning that uses a little amount of labeled data like Supervised Machine Learning and a larger amount of unlabeled data like Unsupervised Machine Learning to train the algorithms. First, the labeled data is used to partially train the Machine Learning Algorithm, and then this partially trained model is used to pseudo-label the rest of the unlabeled data. Finally, the Machine Learning Algorithm is fully trained using a combination of labeled and pseudo-labeled data.

Reinforcement Machine Learning

Well, here are the hypothetical students who learn from their own mistakes over time (that's like life!). So the Reinforcement Machine Learning Algorithms learn optimal actions through trial and error. This means that the algorithm decides the next action by learning behaviors that are based on its current state and that will maximize the reward in the future. This is done using reward feedback that allows the Reinforcement Algorithm to learn which are the best behaviors that lead to maximum reward. This reward feedback is known as a reinforcement signal.

1.3.3 Popular Machine Learning Algorithms

Let's look at some of the popular Machine Learning algorithms that are based on specific types of Machine Learning.

A) Supervised Machine Learning

more popular algorithms in these categories are:

Supervised Machine Learning includes Regression and Classification algorithms. Some of the

Linear Regression Algorithm

The Linear Regression Algorithm provides the relation between an independent and a dependent variable. It demonstrates the impact on the dependent variable when the independent variable is changed in any way. So the independent variable is called the explanatory variable and the dependent variable is called the factor of interest. An example of the Linear Regression Algorithm usage is to analyze the property prices in the area according to the size of the property, number of rooms, etc.

Logistic Regression Algorithm

The Logistic Regression Algorithm deals in discrete values whereas the Linear Regression Algorithm handles predictions in continuous values. This means that Logistic Regression is a better option for binary classification. An event in Logistic Regression is classified as 1 if it

occurs and it is classified as 0 otherwise. Hence, the probability of a particular event occurrence is predicted based on the given predictor variables. An example of the Logistic Regression Algorithm usage is in medicine to predict if a person has malignant breast cancer tumors or not based on the size of the tumors.

Naive Bayes Classifier Algorithm

Naive Bayes Classifier Algorithm is used to classify data texts such as a web page, a document, an email, among other things. This algorithm is based on the Bayes Theorem of Probability and it allocates the element value to a population from one of the categories that are available. An example of the Naive Bayes Classifier Algorithm usage is for Email Spam Filtering. Gmail uses this algorithm to classify an email as Spam or Not Spam.

B) Unsupervised Machine Learning

Unsupervised Machine Learning mainly includes Clustering algorithms. Some of the more popular algorithms in this category are:

K-Means Clustering Algorithm

Let's imagine that you want to search the name "Harry" on Wikipedia. Now, "Harry" can refer to Harry Potter, Prince Harry of England, or any other popular Harry on Wikipedia! So Wikipedia groups the web pages that talk about the same ideas using the K Means Clustering Algorithm (since it is a popular algorithm for cluster analysis). K Means Clustering Algorithm in general uses K number of clusters to operate on a given data set. In this manner, the output contains K clusters with the input data partitioned among the clusters.

1.3.4 Limitations of Machine Learning

Machine Learning Algorithms are trained using data sets. And unfortunately, sometimes the data may be **biased** and so the ML algorithms are not totally objective. This is because the data may include human biases, historical inequalities, or different metrics of judgement based on

gender, race, nationality, sexual orientation, etc. For example, Amazon found out that their Machine Learning based recruiting algorithm was biased against women. This may have occurred as the recruiting algorithm was trained to analyze the candidates' resumes by studying Amazon's response to the resumes that were submitted in the past 10 years. However, the human recruiters who analyzed these resumes in the past were mostly men with an inherent bias against women candidates that were passed on to the AI algorithm.

This means that some Machine Learning Algorithms used in the real world may not be objective due to biased data. However, companies are working on making sure that only objective algorithms are used. One way to do this is to preprocess the data so that the bias is eliminated before the ML algorithm is trained on the data. Another way is to post-process the ML algorithm after it is trained on the data so that it satisfies an arbitrary fairness constant that can be decided beforehand.

- Each narrow application needs to be specially trained
- Require large amounts of hand-crafted, structured training data
- Learning must generally be supervised: Training data must be tagged
- Require lengthy offline/ batch training
- Do not learn incrementally or interactively, in real-time
- Poor transfer learning ability, reusability of modules, and integration
- Systems are opaque, making them very hard to debug
- Performance cannot be audited or guaranteed at the 'long tail'
- They encode correlation, not causation or ontological relationships
- Do not encode entities or spatial relationships between entities
- Only handle very narrow aspects of natural language
- Not well suited for high-level, symbolic reasoning or planning

1.4 WORD CLOUD

Word clouds (also known as text clouds or tag clouds) work in a very simple way that the more a specific word appears in a source of text data (such as a speech, blog, post, or database), the bigger and bolder it appears in the word cloud. A word cloud is a collection, or cluster, of words represented in different sizes. The bigger and bolder the word appears, the more often it is raised up within a given text and the more important it is. Also known as tag clouds or text clouds, these are ideal ways to pull out the most relevant parts of textual data, from blog posts to databases. They can also help business users to compare and contrast two different pieces of text to find the word related similarities between the two.

Fig. 1.3 WORDCLOUD

1.4.1 What is a Word Cloud Generator?

As its name implies, an online word cloud generator is a tool that scans a body of text, turning it into component words.from there, it can create a word cloud that highlights the most frequently mentioned words. If you don't prefer the cluster shape, most tools enable you to format the word cloud in various ways, including:

- Horizontal lines
- Columns
- Formed to fit a certain shape

Most providers will also allow users to choose different layouts, fonts and color schemes depending on their preference. This means you can make one to match the color scheme of your brand, your partners, or your clients. While the color used on a word cloud holds a primarily aesthetic value, you can contrast the hues to help categorize words or illustrate a separate data variable.

1.4.2 Why Use a Word Cloud Generator?

- Search Engine Optimization
- Understanding Client Issues
- Quickening Business Actions
- Analyzing Employee Sentiment
- Simplifying Technical Data
- Searching for Patterns in Data

1.4.3 Where Word Clouds Excel ?

In the right setting, word cloud visualizations are a powerful tool. Here are a few instances when word clouds excel:

Finding Customer Pain Points and Opportunities to Connect -

Do you collect feedback from your customers? (You should!) Analyzing your customer feedback can allow you to see what your customers like most about your business and what they like least. Pain points (such as "wait time," "price," or "convenience") are very easy to identify with text clouds.

Understanding How Your Employees Feel About Your Company -

Text cloud visualization can turn employee feedback from a pile of information you'll read through later to an immediately valuable company feedback that positively drives company culture.

Identifying New SEO Terms to Target -

In addition to normal keyword research techniques, using a word cloud may make you aware of potential keywords to target that your site content already uses.

1.4.4 Limitations of Word Cloud –

- They do not capture words that mean the same thing.
- They do not capture words that mean the same thing.
- They lack context.
- They abscure the relative importance of themes.
- They are prone to bias.

1.5 PRODUCT RECOMMENDATION SYSTEM

A product recommendation system is a solution that provides relevant product suggestions to the customers in real time. It is a powerful data filtering platform that depends on algorithms, artificial intelligence, machine learning, and other data analyzing practices. It is a concatenation collecting, storing, analyzing, and filtering customers' data to provide highly personalized relevant products to each and every customer. Relevant products meet the customers'

requirements, tastes, and preferences. The quality of data should be very high to achieve such refined targeting at an individual level. But most importantly, we need the right tool to understand the customer data and business needs. Successful product recommendations can increase customer engagement, conversion rates, and revenue. But, you must get the right product recommendation system for your sector and business to maximize its potential.

1.5.1 Benefits of Recommendation

- Improved user experience
- Improved e-commerce conversion rate
- Higher retention rate
- Reduced online shopping cart abandonment
- Optimized inventory management
- Improved customer loyalty
- Improved time and cost efficiency
- Increase in revenue
- Improved user engagement
- Increased sales

1.5.2 Machine Learning Techniques

Now, when you know the basics, let us explain to you what machine learning techniques can be used in these systems.

Content-based Filtering

Such a system defines what products users click and buy, what pages they view, etc. Then, on the basis of this information, a user profile is created. Compared to the product list, the profile is used to provide recommendations.

Collaborative Filtering

This technique implies using recommendations from users. Their behavior and preferences are analyzed and then used to determine similarities between users. As a result, you will understand which products a particular user may like thanks to their similarity to other customers.

Complementary Filtering

The complementary filtering technique analyzes the probability of a few products being purchased together. In other words, it defines complementary products of a specific item. When a user buys this item, they get a recommendation to purchase a complementary one.

Demographic-based Filtering

Clearly, this technique uses demographic information of the customers. It recommends products bought by users with a similar demographic profile.

Hybrid Recommendation System

There is no need to focus on a single technique. To get better results and more accurate suggestions, you can combine several of them in a hybrid system.

1.5.3 Limitations of Recommendation

The first challenge you may face is processing huge data sets to get real-time predictions. AI Consulting is a great help, but you will still have to set up the parameters. The larger the data set is, the harder it will be to reach the maximum accuracy. Use large-scale assessment methods to overcome the task.

The second issue is that your system will have no information about new users. Thus, it won't be able to recommend something to them on the basis of their profiles and preferences. To solve this problem, you can recommend popular products or use contextual information (for instance, the user's location). New products also have no reviews or clicks for some time, until users discover

them. You can recommend such products using their metadata and the content-based filtering technique.

And, finally, the diversity. Collaborative methods are effective, but sometimes they can't deliver sufficient diversity. To deal with this issue, you can recommend products disliked by people who are not similar to a specific user.

The Several recommendation systems have been proposed that are based on collaborative filtering, content and hybrid recommendation methods but these have some problems which are the challenges for research work. It is required to work on this research area to explore and provide new methods that can reduce the challenges and provide recommendation in collaborating filtering a wide range of applications while considering the quality and privacy aspects. Thus, the current recommendation system needs improvement for present and future requirements of better recommendation qualities.

1.6 DISSERTATION OUTLINE

The remaining part of the dissertation is organized as follows:

Chapter 2

This chapter gives a summary of the literature in the field of Sentiment Analysis. Literature review was done in order to have clear understanding of the topic, the problem statement and the progress done so far.

Chapter 3

This chapter introduces the methodology used in our research. It also talks about the purpose and the objectives of the research.

Chapter 4

This chapter briefs about all the processes involved in research. Model approach is been presented with the help of flowcharts and diagrams.

Chapter 5

In this chapter, the experimental results and analysis are discussed. It also presents all the results of the research.

Chapter 6

In this chapter conclusion and some of the future scopes of this work are discussed.

CHAPTER 2

LITERATURE REVIEW

2.1 LITERATURE SURVEY

Levent Guner from KTH Royal Institute of Technology, Stockholm selected 60,000 random product reviews from Amazon. He used the dataset available in Kaggle that contains 4 million reviews. [1] The performance was compared with three different algorithms namely Naïve Bayes (NB), Support Vector Machine (SVM) and Long short-term memory network (LSTM). The authors used multiple performance metrics to determine the best performing classification algorithm on the test set. The performance metrics used were Accuracy, Area Under Curve (AUC), Precision, Recall and F1- score. Based on the results of the evaluation, their study concluded that the LSTM model performed the best with precision > 0.90 and AUC = 0.96 for binary classification (positive and negative).

Xing Fang and **Justin Zhan** collected over 5.1 million product reviews in 4 key categories: beauty, book, electronics, and home. [2] They analyzed these reviews with 3 different classifiers, namely, Naïve Bayes, Support Vector Machine and Random Forest. Their paper addressed the basic question of evaluating sentiments, categorizing sentiment polarity and concluded with random forest generating more reliable results. As per their findings, for larger data sets SVM worked better than Naïve Bayes.

Wan Liang Tan performed both traditional machine learning algorithms including Naïve Bayes, SVM, K-Nearest Neighbor and Deep Learning Network Models such as Recurrent Network Models and LSTM on Amazon reviews dataset. [3] They collected 34627 reviews and divided it into 21000 and 13627 records of training and test datasets respectively. In terms of test accuracy, LSTM performed best among all of them with 71.5% accuracy. One of the key reasons for not high enough accuracy was the imbalance in their data, as they concluded.

Callen Rain used Naïve Bayes and Decision-List classifiers to classify product reviews (category: books) from Amazon as positive and negative. [4] He used a corpus that includes

50,000 reviews of 15 items that serve as the research dataset. The features such as bag-of-words and bigrams are compared with each other in their usefulness in labelling positive and negative reviews correctly. His analysis showed that Naive Bayes performed better than the decision-list and bag of words ended up being the best form of feature extraction.

Nishit Shrestha and **Fatma Nasoz** analyzed the opinions of Amazon.com reviews. They developed a model using Recurrent Neural Networks (RNN) with Gated Recurrent Unit (GRU) that learned low dimensional review vector representation using paragraph vectors and product embedding. [5] The data used in this analysis is a collection of about 3.5 million product reviews gathered from Amazon.com. Paragraph Vectors (PV) are very much inspired by word vectors. PV system learns vectors by predicting the next term, given several sampled contexts from a paragraph. The concatenation of review embedding developed from paragraph vectors and GRU-derived product embedding is used to train a Support Vector Machine (SVM) to classify sentiments. With only review embedding, the anticipated classifier provided 81.29 percent accuracy. The product embedding inclusion improved the accuracy to 81.82 percent. Authors believe that a similar technique can be used to learn user information.

In a **research article** different approach has been implemented for sentimental analysis in this research an algorithm called a BoW (Bag of words) is used in which the relationship between the words was not considered. [6] To measure the sentiment for the whole sentence, the sentiment of every single word of the sentences has been individually determined and values are collected using some aggregation function. Along with this opinion summarization method based on features driven can be used. For each product a specific feature and their attributes are obtained, and the general feature for each product class is obtained. Then polarity is assigned to each function with the aid of Sequential Minimal Optimization and Support Vector Machines.

There have been several academic papers published so far on product ratings, sentiment analysis, and opinion mining. On Yelp's ranking dataset, for example, **Xu Yun** from Stanford University used existing supervised learning algorithms like the perceptron algorithm, naive bayes, and supporting vector machine to predict a review's rank. [7] They carried out cross validation with 70% of the data.

The opinionated reviews also contain other information that can be used to ascertain the sentiment about a product. **Venkata Rajeev P** uses the reviews from flipkart.com and proposes the combination of four parameters: star ratings of the product, the polarity of the review, age of review and helpfulness score, for determining the opinion of a product. [8]

Shoiab Ahmed proposes that the count of scored opinion words be classified into seven possible categories i.e. strong-positive, positive, weak-positive, neutral, weak-negative, negative, strong-negative. [9] Sentiment analysis is then done with the help of these score counts.

D V Nagarjuna Devi proposes a system that uses a supervised classification approach called as support vector machine [10]. This paper claims that the proposed classifier approach gives out the best result. It also identifies various challenges in sentiment analysis like sarcasm and conditional sentences, grammatical errors, spam detection and anaphora resolution. sentence level classification is done on input data which is further classified according to the subjectivity/objectivity. Further aspect extraction is done using SentiWordNet. This is then further fed to SVM classifier to find the overall opinion.

Maria Soledad Elli did sentiment from considering reviews of customer. They have analyzed result to develop model for business. [11] Author presented that tool is providing better accuracy. Research have make use of Multinomial Naive Bayesian. It is acting as classifiers. Mechanism is also supporting vector machine.

The task of mining the features is of particular importance and many methods are suggested for it. [12] **Weishu Hu** divided the opinion analysis tasks into three steps: identifying the opinion sentences and their polarity, mining the features that are commented upon by customers, and removing incorrect features.

Product review sentiment analysis, also called as opinion mining, is a method of ascertaining the customers' sentiment about a product on the basis of their reviews. [13] **Liu** classifies the opinion mining tasks into three levels: document level, sentence level and phrase level.

The primary focus of product review system is identifying the adjective word in a sentence and identifying the sentiment behind it. [14] **Yan Luo** suggests the final sentiment score of the review to be the cumulative sentiment score of all the adjectives in that review

Ronan Collobert has made use of convolutional network. [15] Semantic role labeling task has been performed with the objective of avoiding too much operation oriented engineering of characteristics.

On the other hand, in a paper, the author **R Socher** proposed using recursive neural networks to achieve a better understanding compositionality in tasks such as sentiment detection. [16] In this paper, we want to apply both traditional algorithms including Naive Bayesian, K-nearest neighbor, Supporting Vector Machine and deep-learning tricks.

2.2 COMPARTIVE LITERATURE REVIEW

Table 2.1 Comparative Literature Review

S No.	Author / Year	Objective of Research	Methodology	Limitations
1.	Levent Guner 2020	To test the accuracy of machine learning algorithms	Machine Learning	Only accuracy was tested, no future scope
2.	Xing Fang and Justin Zhan 2018	To test sentiments	Sentiment Analysis and Machine Learning	Sarcasm was misinterpreted
3.	Wan Liang Tan 2020	Performance evaluation of different techniques	Sentiment classification	Only performance was tested, no future scope
4.	Callen Rain 2013	Implementing Sentiment analysis in Amazon reviews using probabilistic machine learning	Machine learning	Research is providing solution on the basis of probability that leads to degradation in accuracy
5.	Nishit Shrestha and Fatma Nasoz 2020	Classifying reviews for future decisions	Natural Language Processing	Old dataset was taken rather than current one
6.	Xu Yan 2015	Implementing Sentiment analysis of yelps ratings based on text reviews	Sentiment Analysis	Research is not provided wide scope
7.	Venkata Rajeev P	Sentiment analysis of movie reviews	Sentiment Analysis and Machine Learning	Sentiments were only classified as positive and negative
8.	Shoiab Ahmad 2018	Proposing KNN classifier based approach for multiclass sentiment	KNN classifier	This work is suffering from performance issues

		analysis		
9.	D V Nagarjuna Devi 2003	Performing Opinion extraction and semantic classification of product reviews	Semantic Classification	Research has not considered optimized solution
10.	Maria Soledad Elli 2017	To propose Amazon reviews, business analytics with sentiment analysis	Analyzing the Sentiment	There is lack of accuracy is prediction
11.	Weishu Hu 2015	Implementing Sentiment analysis of yelps ratings based on text reviews	Sentiment Analysis	Research has not provided wide scope
12.	B Lui 2012	To perform Opinion Mining	Sentiment Analysis	There is lack of accuracy and flexibility
13.	Yan Luo	Methods' performance Comparison	Machine Learning	Misspellings and grammatical mistakes may cause the analysis to overlook important words or usage
14.	Ronan Collobert 2011	To implement Natural language processing from scratch.	Machine learning	Research failed to provide solution for semantic review
15.	R Socher 2013	Presenting Recursive deep models for semantic compositionality over a sentiment treebank.	Recursive Deep Model	Recursive deep model wastes lot of time during training

2.3 CONCLUSION

The purpose of performing the literature review is to highlight the researches done in the past by other researchers in the field of extracting sentiments of products through Sentiment Analysis and Machine Learning. Yet, there is large scope of improving accuracy of work, as improvements in machine learning models can be seen with the passage of time.

CHAPTER-3

METHODOLOGY

3.1 PROPOSED METHODOLOGY

The dataset used for this project is from the amazon.com. The reviews in the dataset are consists of the attributes such as: Reviewer ID, Product ID, Review Text, Rating and time of the review. The main source of data used is the product reviews from Amazon. The reviews of a **JBL Digital Soundbar** have been obtained by building a web crawler. The web crawler has been written in Python using a scrapping library. Along with the review text, some additional data related to the reviews such as reviewer name, review date, overall rating and comments were also obtained. The crawler is called periodically to get the most up-to-date reviews. Each review is generally treated as a sentence or a group of sentences. They are cleaned and stored in a .CSV file on google sheet. The first stage of analysis involves preprocessing of the reviews. Preprocessing involves the following operations: stemming, stopword removal and part-of-speech tagging.

Then, sentiment analysis is performed on the preprocessed reviews and overall sentiment score for each review is generated. Further for feature extraction, there are two cases:

Single Feature – If the review contains only a single feature, then the sentiment score of the review is assigned to the feature.

Multiple Features – Some reviews have multiple features contained in them. So the above procedure will not work in this case. Rules are defined to extract multiple features and assign the correct sentiment score to those features. For reviews containing more than one sentence, first check if the review contains a word from an adjective word set or not. If it does not contain one, then it is assumed to be objective. If it does contain an adjective, the feature that corresponds to that adjective is found by looking for a set of predetermined nouns near that adjective.

The plan here is to implement SVM, Naïve Baye and Logistic Regression. Sentiment scores range from 1 to 5, 1 being the most negative, 5 being the most positive and 3 being neutral. These scores are then averaged for each feature and soundbar and stored in a database. The

recommendation engine takes in a set of user preferences in terms of the features that the user would like in the product. It presents the most suitable product to the user based on the scores assigned to each sound bar in the previous step.

3.2 PROBLEM STATEMENT

An application that collects reviews from the users about a certain product and analyzes them. It would segregate the reviews into positive and negative reviews. The negative reviews will be helpful to the companies to further enhance their product based on the user's feedback. The application will provide the pros and cons of the individual feature of the product and hence, reports about the sentiment analysis performed on the products. Then the further aim is to create a recommendation system that recommends products to users according to the feature requirement of user.

There is a strong correlation between user reviews on Amazon, it is very costly to obtain sentiment labels for large training data and expressions as data of Amazon is unstructured, informal and fast evolving. Amazon has huge data that is present in both structured and unstructured formats. A dataset will be taken and then classified according to its sentiment. It is to collect valuable reviews of a product.

The process of separating emotions, comments from the reviews will begin with feature extraction.The process of labeling begins where the words present in reviews are classified as per five categories, i.e, **Highly Recommended, Recommended, Risky, Highly Risky, Not Recommended.**

3.3 OBJECTIVE

- By using this work
- A customer can get receive recommendation for the product
- A business organization can get the recommendation whether to keep or to discard it in future.

3.4 MODEL ARCHITECTURE

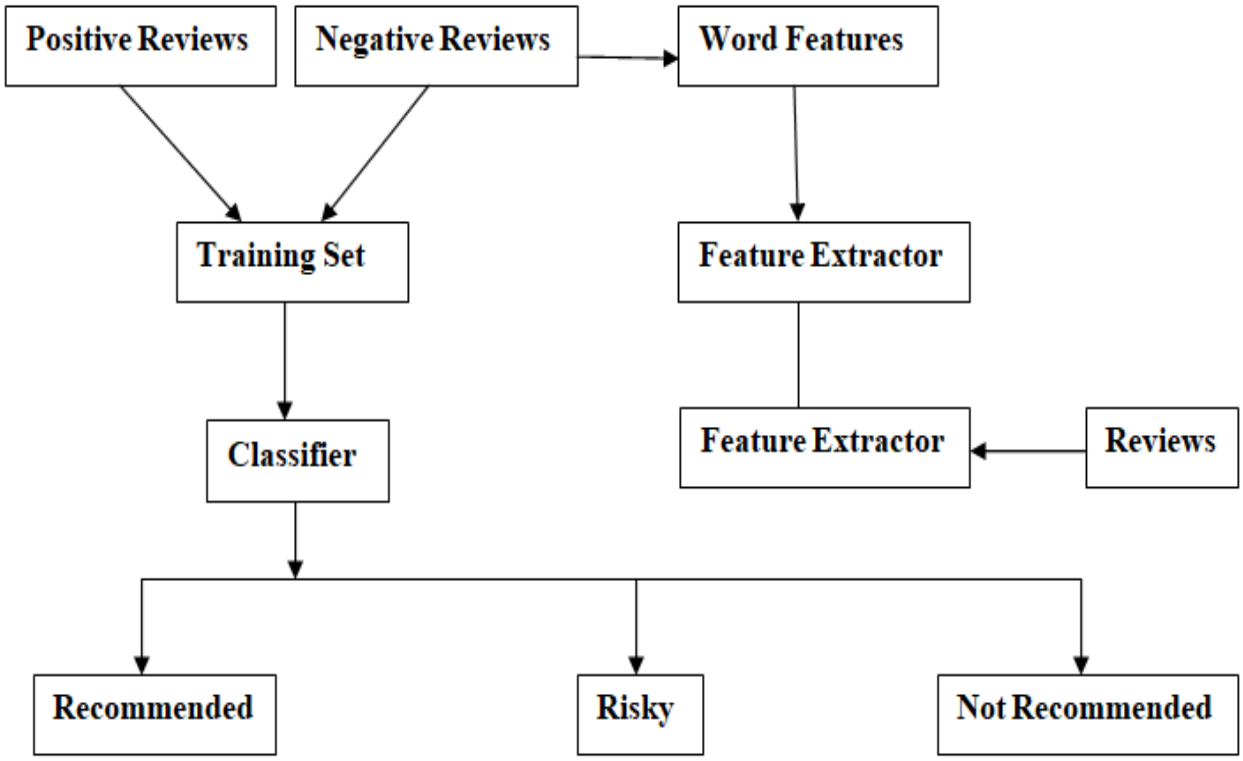


Fig. 3.1 Sentiment Analysis Procedure Flowchart

3.4.1 Data Collection

The very first job in the process of sentiment analysis is data collection. Data can be collected from various sources like any website, from the several online opinion sets & ratings.

Some methods of data collection are –

- Forms and Questionnaires

- Online Tracking
- Documents and Records
- Interviews
- Social Media Monitoring

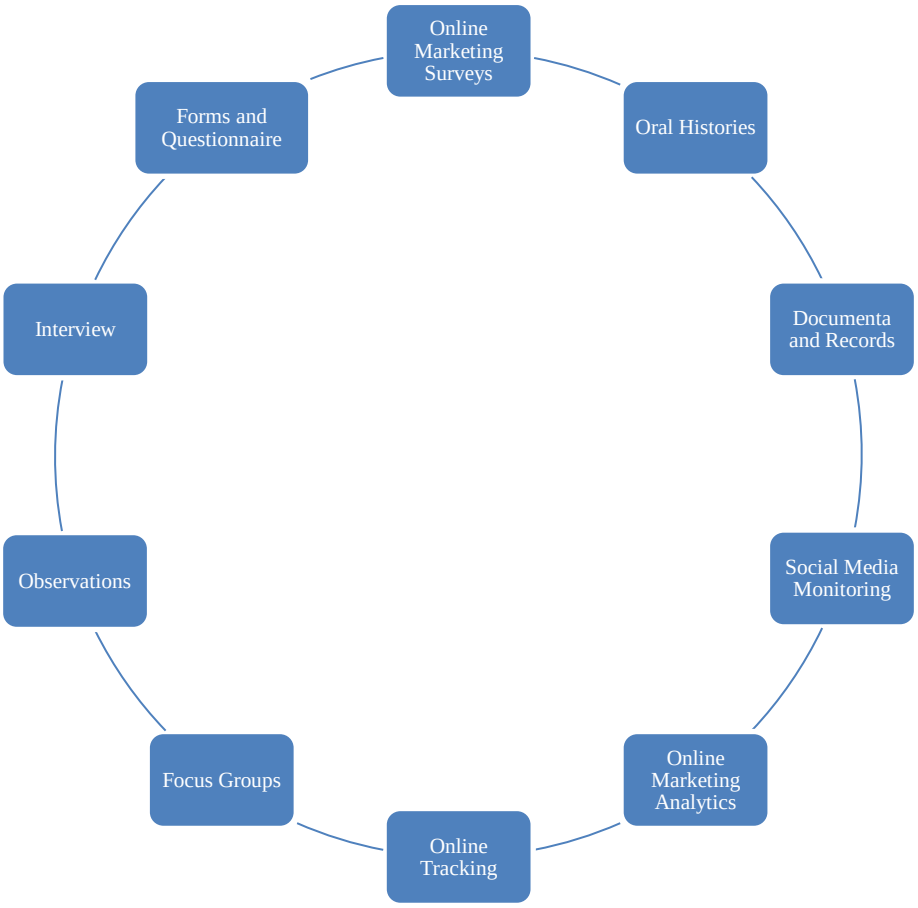


Fig. 3.2 Data Collection Methods

3.4.2 Data Preprocessing

It is the cleaning process of data. Unrequired words & symbols are omitted. This is required for further processing to be streamlined. Part of this move is eliminating hyperlinks, repeated sentences, emoticons, and special characters. It also performs lemmatization and stemming. Finally, it takes a reduced collection of features and passes them to the classifiers.

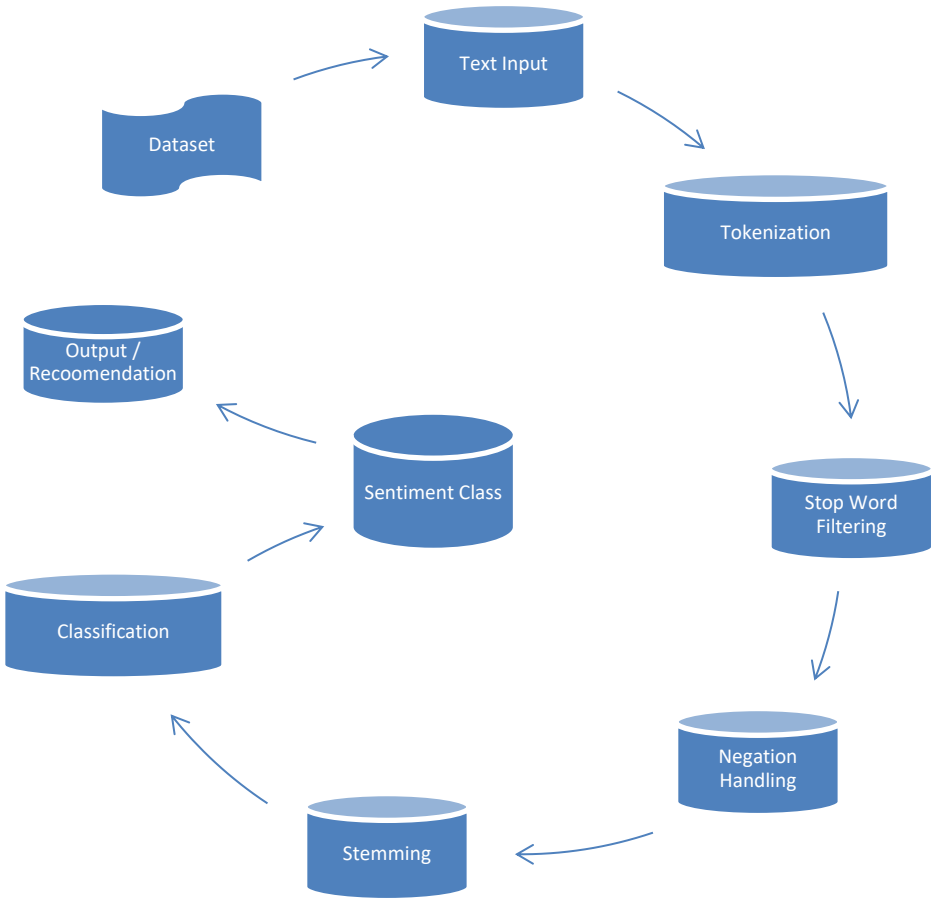


Fig. 3.3 Data Preprocessing Procedure

Tokenization- Token is defined as the minimal unit that a machine understands and processes at a time. All the text strings are processed only after they have undergone tokenization, which is the process of splitting the raw strings into meaningful tokens.

Lemmatization- Lemmatization is a methodical way of converting all the grammatical/inflected forms of the root of the word. Lemmatization makes use of the context and POS tag to determine the inflected form (shortened version) of the word and various normalization rules are applied for each POS tag to get the root word (lemma).

Stemming- Stemming is the process of obtaining the root word from the word given. Using efficient and well-generalized rules, all tokens can be cut down to obtain the root word, also known as the stem. Stemming is a purely rule-based process through which we club together

variations of the token. For example, the word sit will have variations like sitting and sat. It does not make sense to differentiate between sit and sat in many applications, thus we use stemming to club both grammatical variances to the root of the word. Stemming is in use for its simplicity. But in the case of dravidian languages with many more alphabets, and thus many more permutations of words possible, the possibility of the stemmer identifying all the rules is very low. In such cases we use the lemmatization instead. Lemmatization is a robust, efficient and methodical way of combining grammatical variations to the root of a word.

Stop Word removal- Stop words are the most commonly occurring words, that seldom add weightage and meaning to the sentences. They act as bridges and their job is to ensure that sentences are grammatically correct. It is one of the most commonly used pre-processing steps across various NLP applications. Thus, removing the words that occur commonly in the corpus is the definition of stop-word removal.

rules is very low. In such cases we use the lemmatization instead. Lemmatization is a robust, efficient and methodical way of combining grammatical variations to the root of a word.

3.4.3 Data Classification

The most critical aspect of a system for sentiment analysis is a classifier. Classification is achieved in negatives, positive, or neutrals categories. A third of the database is usually used as training sets to generate the classifiers. To a large degree, the precision of the classifier relies on the training collection. By using machine learning classifiers like SVM, Bayesian Classifiers and so on, the classification can be performed. However, before training and testing the classifier, machine learning classifiers do feature extraction, which can also use deep neural networks for classifying the data.

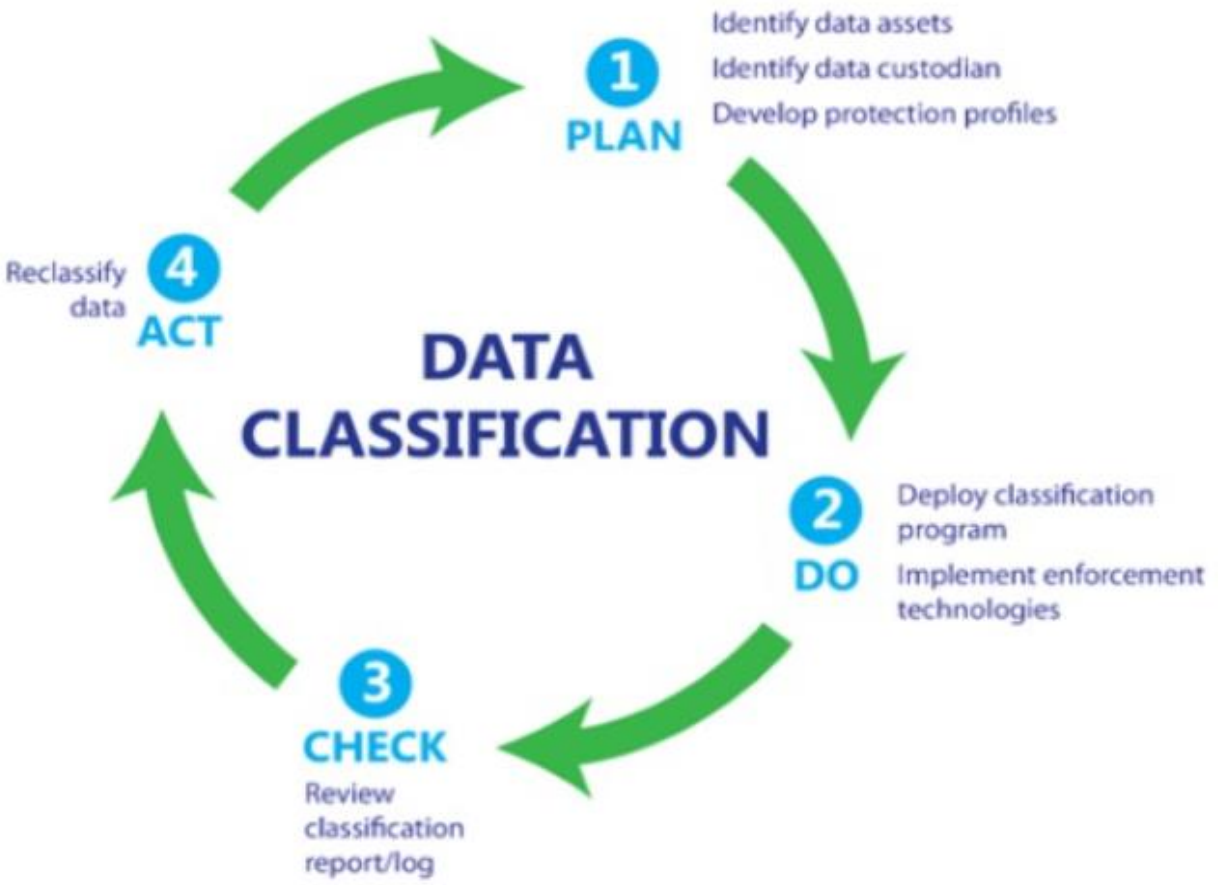


Fig. 3.4 Data Classification Method

3.4.4 Data Exploration

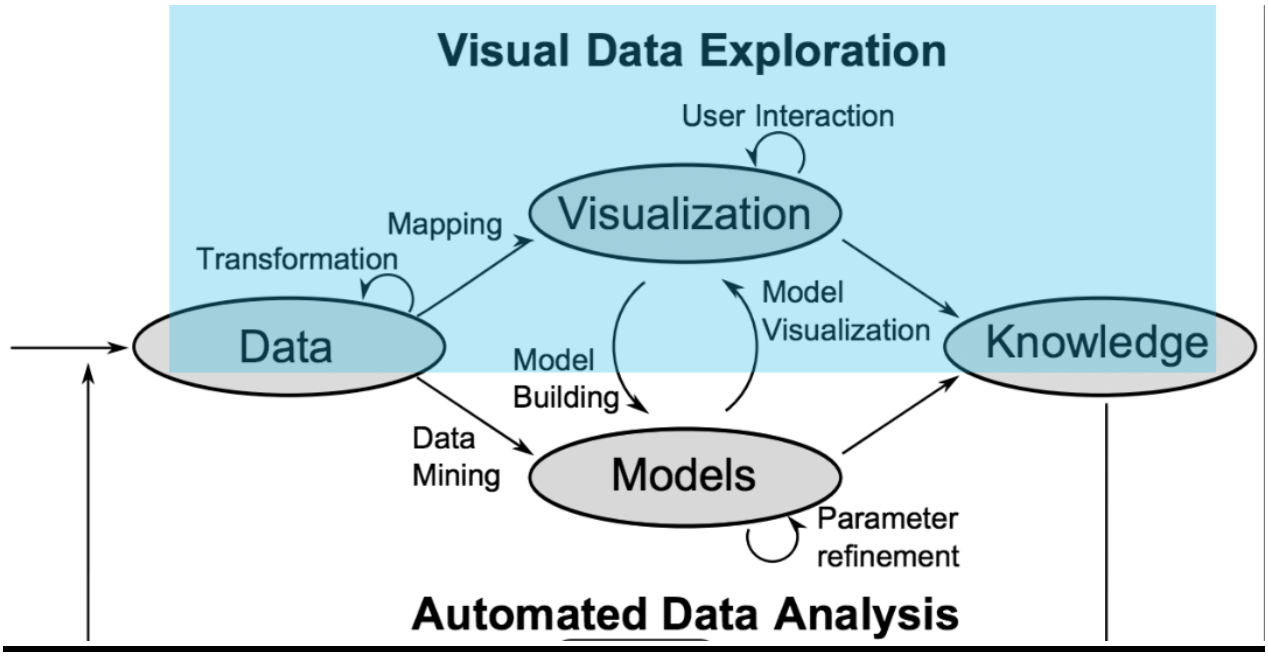


Fig. 3.5 Data Exploration Method

Naive Bayesian Classifier

It is believed that the Naive Bayesian Classifier is very simple and easy in terms of implementation. This is not any single algorithm but consists of the set based on Bayes theorem of various classification algorithms. A term used to define an event's probability. This probabilistic classifier utilizes and analyses all the characteristics present in the vector of the function differently, i.e., it considers them independently of each other.

By analyzing a pre-categorized collection of documents, we can learn the pattern. This model states that the conditional probabilities of the event P(A) occurring could be determined in presence of the two events, P(A) & P(B) if P(B) has already occurred.. The Naive Bayes classifier's input for training consists of preprocessed data along with its extracted features. The classifications process is conducted on the data set of test data after completing the training and then, depending on the outcomes, the new data. A polarity of the feelings of the data is given by

this classification method. For instance, the "It was good" review statement would have resulted in positive polarity.

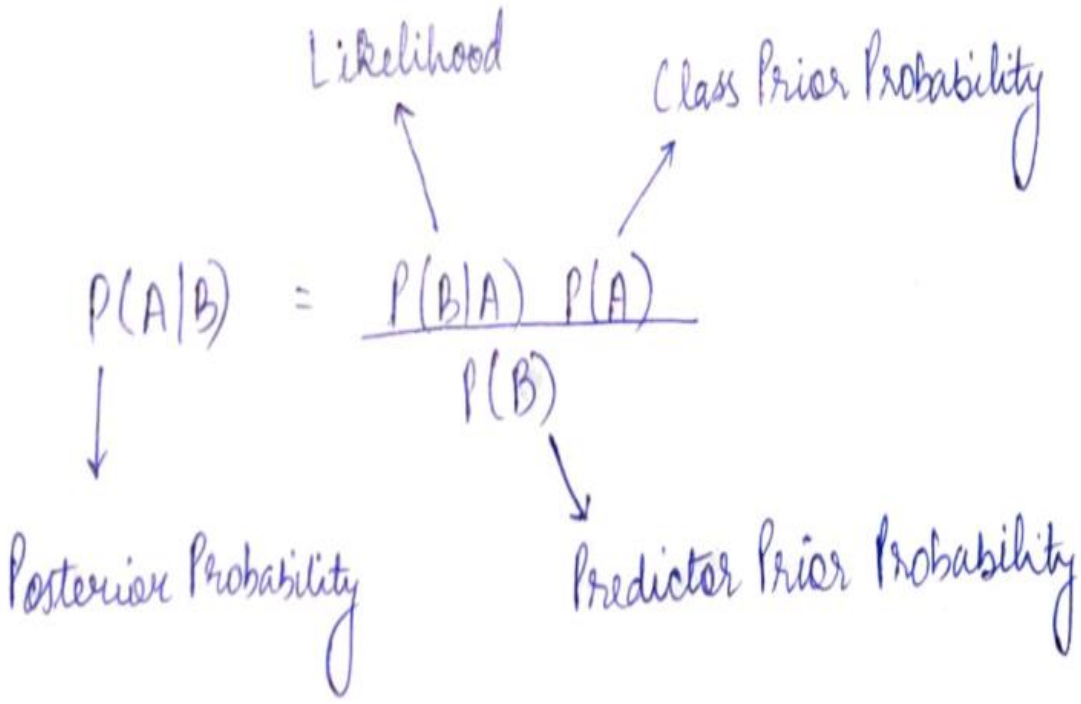


Fig. 3.6 Formula for Naive Bayesian Classifier

SVM Classifier

SVM is a popular machine learning technique that employs a statistical approach. It is extremely effective at text classification. There is an n-dimensional space in the SVM, in which n represents number or quantity of features presented in a vector of the function. In the n-dimensional space, each of the data elements presents in the training dataset is registered, the value of each character is the coordinate value. In this particular n-dimensional space, the key concept of this approach is to find linear separators that best differentiate the various groups. SVM uses a differentiation function with the following parameters: "X" is the vector of the function; the weight vector is "w", and the bias vector is "b". On the training set, the weights & preload vector are automatically learned. Between these two classes, a margin that is far from a document is described. The

classifier margins are defined by this distance and indecisive choices are reduced by maximizing this margin. While some features are important to this system, due to the sparse nature of the text, they are correlated and therefore well suited for SVM text classification.

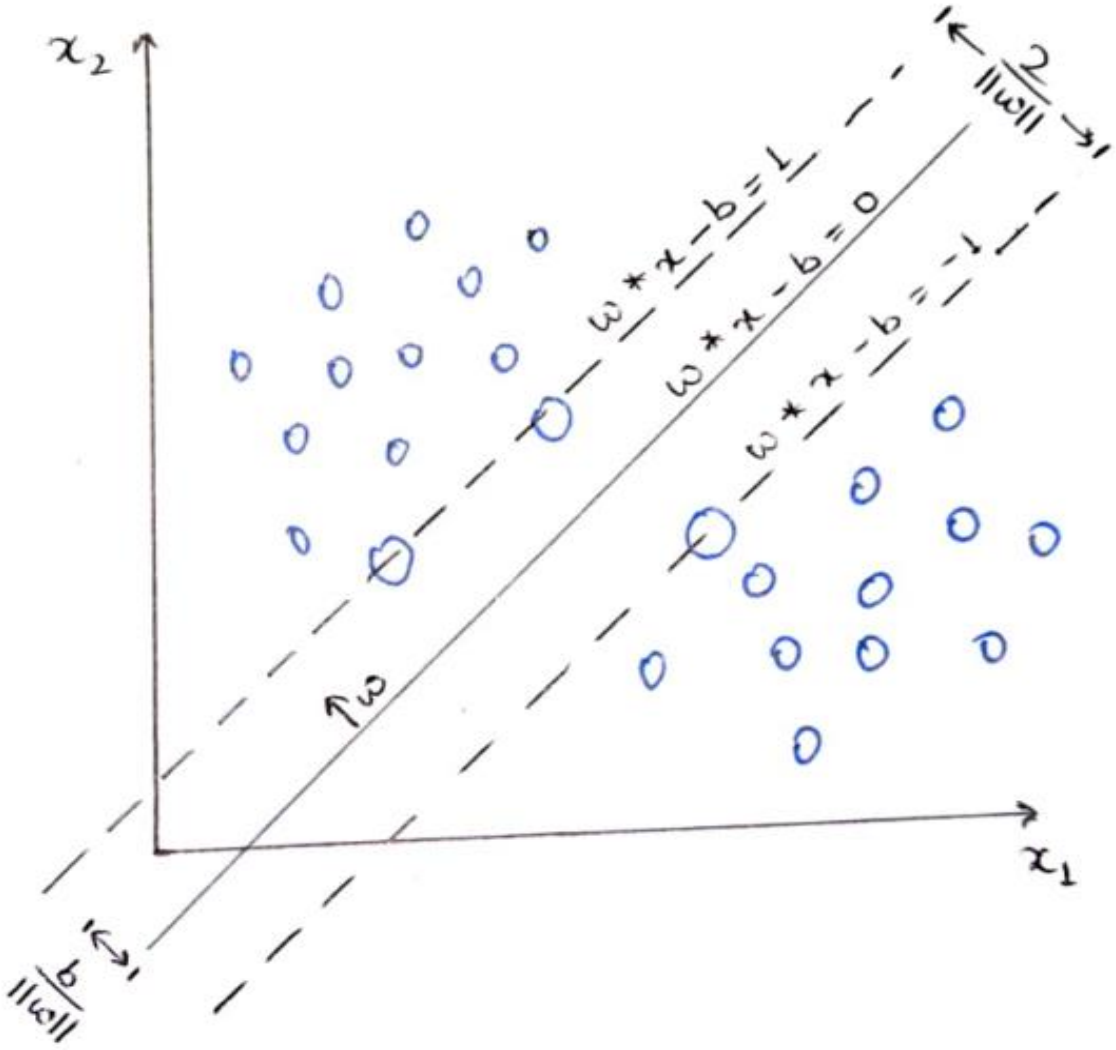


Fig. 3.7 Graph for SVM Classifier

Logistic Regression

A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. For example, a logistic regression could be used to predict whether a political candidate will win or lose an election or whether a high school

student will be admitted or not to a particular college. These binary outcomes allow straightforward decisions between two alternatives. Logistic regression has become an important tool in the discipline of machine learning. It allows algorithms used in machine learning applications to classify incoming data based on historical data. As additional relevant data comes in, the algorithms get better at predicting classifications within data sets. A logistic regression model can take into consideration multiple input criteria.

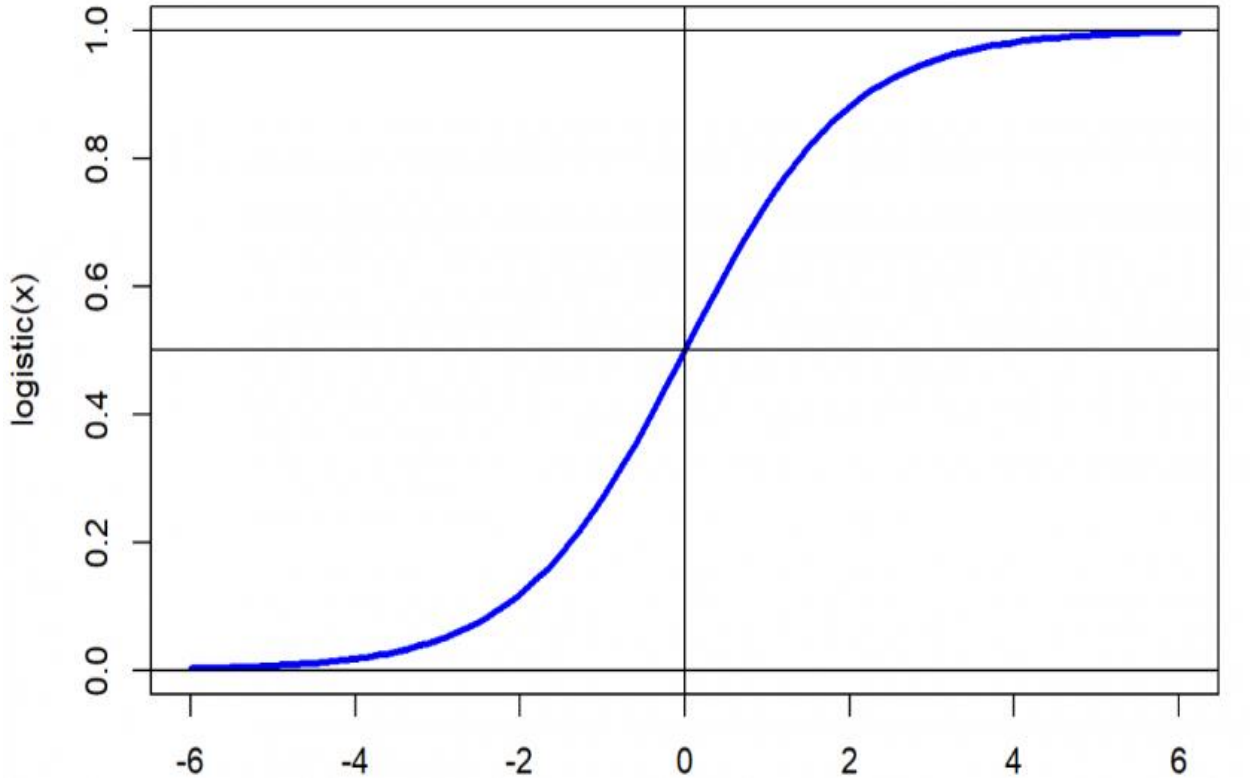


Fig. 3.8 Graph for Logistic Regression

3.4.5 PRODUCT RECOMMENDATION

Product recommendation is done on the basis of polarity of reviews. A customer places a review with respect to the quality, quantity, value, price and experience with the product, so, in the accordance with the previous reviews of the product, the product will further be recommended to the customer if it has high positive polarity, otherwise, it will not be recommended.

A particular product is recommended if-

Table 3.1 Table for Recommendation of Product

S NO.	POLARITY	RECOMMENDATION
1	Positive	Recommended
2	Neutral	Risky
3	Negative	Not Recommended

3.5 TOOLS / APIs USED

The following Tools and APIs were used-

3.5.1 Python

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting

Libraries Imported In Python

- Numpy
- Pandas
- Matplotlib.pyplot
- Re
- String
- Math
- Spacy

- Sklearn.feature_extraction.text
- Textwrap
- Textblob

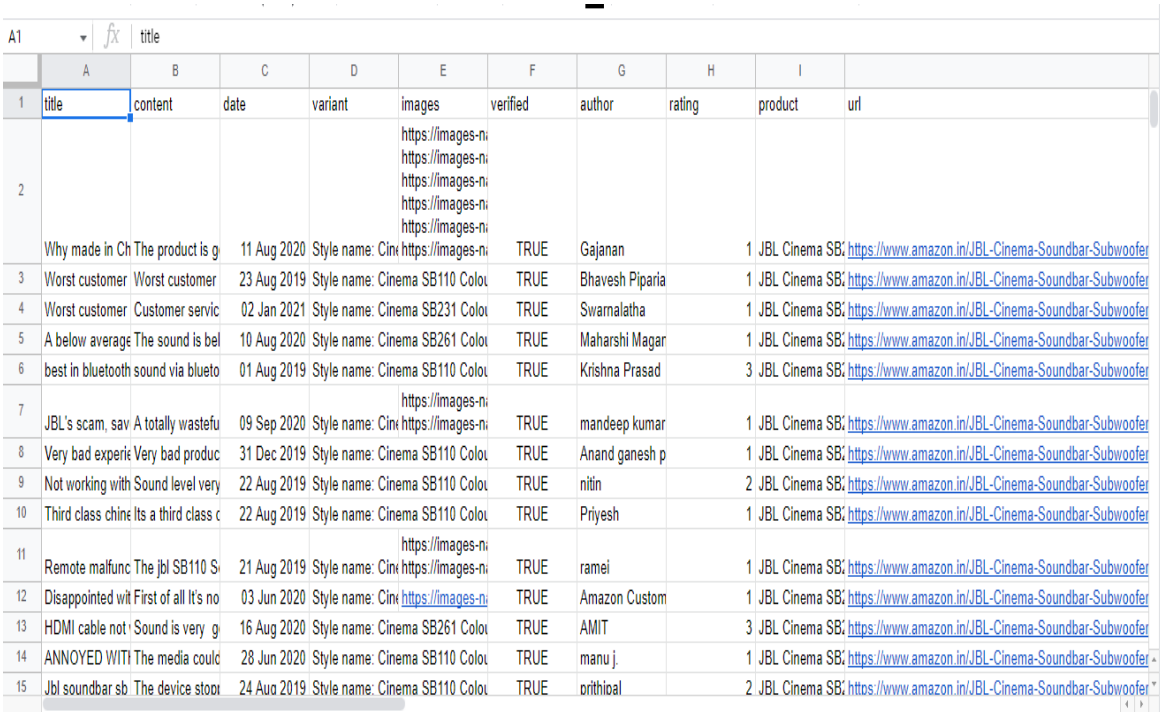
```
# import required libraries

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import re
import string
import math
import spacy
from sklearn.feature_extraction.text import CountVectorizer
from wordcloud import WordCloud
from textwrap import wrap
from textblob import TextBlob
```

Fig. 3.9 Imported Libraries

3.5.2 Google Sheet

Google Sheets is an online spreadsheet app that lets you create and format spreadsheets and work with other people. Click New. This will create and open your new spreadsheet.



	A	B	C	D	E	F	G	H	I	
1	title	content	date	variant	images	verified	author	rating	product	url
2	Why made in Ch	The product is g	11 Aug 2020	Style name: Cin	https://images-ni https://images-ni https://images-ni https://images-ni https://images-ni https://images-ni	TRUE	Gajanan	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
3	Worst customer	Worst customer	23 Aug 2019	Style name: Cinema SB110 Colou		TRUE	Bhavesh Piparia	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
4	Worst customer	Customer servic	02 Jan 2021	Style name: Cinema SB231 Colou		TRUE	Swarnalatha	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
5	A below average	The sound is bel	10 Aug 2020	Style name: Cinema SB261 Colou		TRUE	Maharshi Magar	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
6	best in bluetooth	sound via blueto	01 Aug 2019	Style name: Cinema SB110 Colou		TRUE	Krishna Prasad	3	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
7	JBL's scam, sav	A totally wastefu	09 Sep 2020	Style name: Cin	https://images-ni https://images-ni	TRUE	mandeep kumar	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
8	Very bad experie	Very bad produc	31 Dec 2019	Style name: Cinema SB110 Colou		TRUE	Anand ganesh p	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
9	Not working with	Sound level very	22 Aug 2019	Style name: Cinema SB110 Colou		TRUE	nitin	2	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
10	Third class chine	Its a third class c	22 Aug 2019	Style name: Cinema SB110 Colou		TRUE	Priyesh	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
11	Remote malfunc	The jbl SB110 S	21 Aug 2019	Style name: Cin	https://images-ni https://images-ni	TRUE	ramei	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
12	Disappointed wit	First of all It's no	03 Jun 2020	Style name: Cin	https://images-ni	TRUE	Amazon Custom	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
13	HDMI cable not	Sound is very g	16 Aug 2020	Style name: Cinema SB261 Colou		TRUE	AMIT	3	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
14	ANNOYED WITI	The media could	28 Jun 2020	Style name: Cinema SB110 Colou		TRUE	manu j.	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
15	Jbl soundbar sb	The device stopp	24 Aug 2019	Style name: Cinema SB110 Colou		TRUE	prithipal	2	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer

Fig. 3.10 Google Sheet

3.5.3 MATPLOTLIB

Matplotlib is a **plotting library** available for the Python programming language as a component of NumPy, a big data numerical handling resource. Matplotlib uses an object oriented API to embed plots in Python applications. One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

3.5.4 NLTK

The **Natural Language Toolkit** (NLTK) is a platform used for building Python programs that work with human language data for applying in statistical natural language processing (NLP). It contains text processing libraries for tokenization, parsing, classification, stemming, tagging and

semantic reasoning. Nltk word_ tokenize is extremely important for pattern recognition and are used as a starting point for stemming and lemmatization. Nltk word_tokenize is used to extract tokens from a string of characters using the word tokenize method. It actually returns a single word's syllables.

CHAPTER-4

PROPOSED WORK

4.1 PROPOSED GOAL

- Gather significant reviews from the site.
- Perform feature extraction.
- Classify words present in the Reviews.

An application that collects reviews from the users about a certain product and analyzes them. It would segregate the reviews into positive and negative reviews. The negative reviews will be helpful to the companies to further enhance their product based on the user's' feedback. The application further provides the pros and cons of the individual feature of the product. The application will further provide reports about the sentiment analysis performed on the products. We further aim to create a recommendation system that recommends products to users according to the feature requirement of user.

There is a strong correlation between user reviews on Amazon, it is very costly to obtain sentiment labels for large training data and expressions as data of Amazon is unstructured, informal and fast evolving. Amazon has huge data that is present in both structured and unstructured formats. We will take a dataset and will classify according to its sentiment. It is to collect valuable reviews of a product.

The process of separating emotions, comments from the reviews will begin with feature extraction. The process of labeling begins where the words present in reviews are classified as per five categories, i.e, Recommended, Risky, Not Recommended.

4.2 APPROACH FLOW DIAGRAM

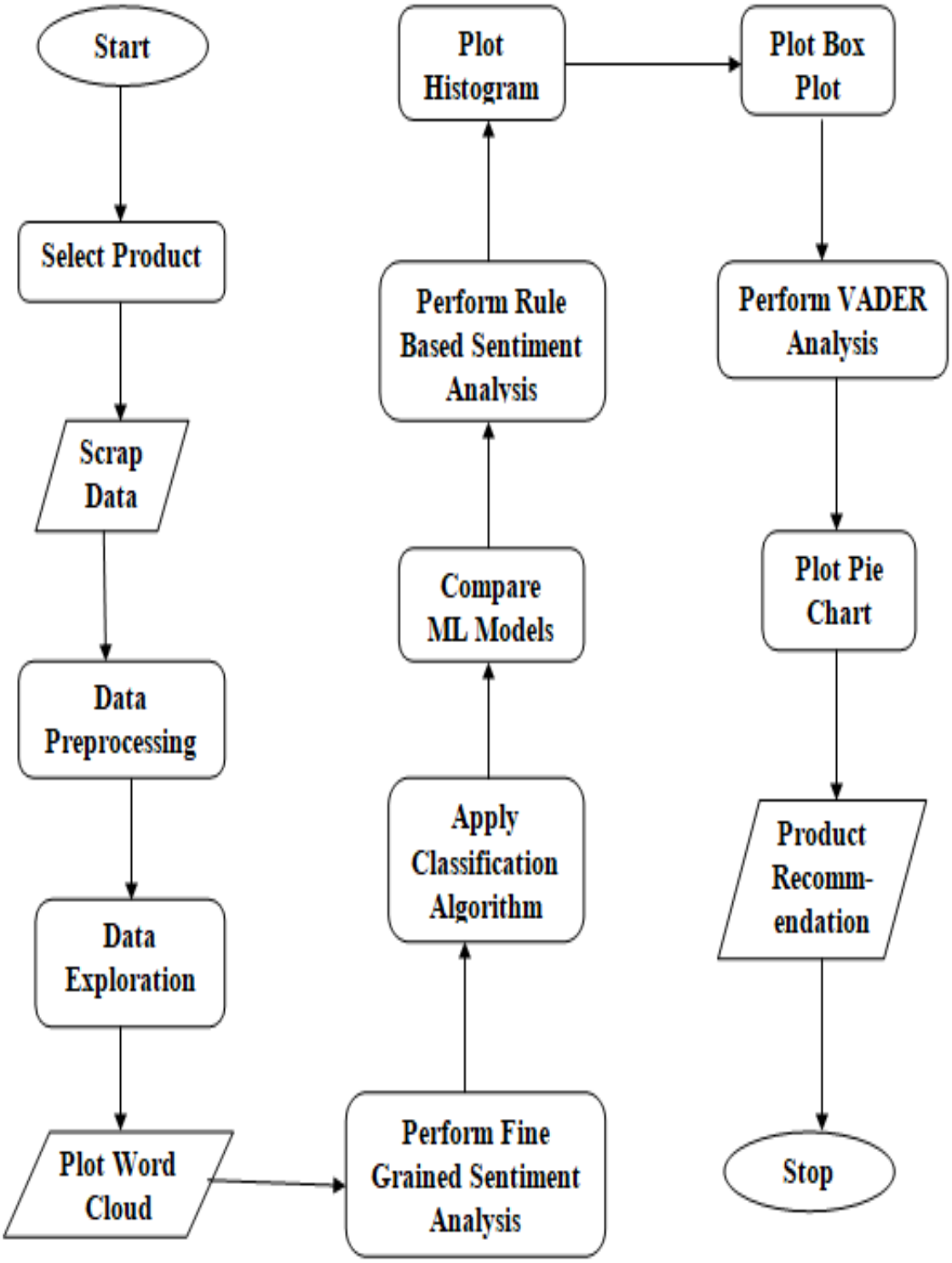


Fig. 4.1 Approach Flowchart

4.3 EXTRACTION OF REVIEWS

I17 | JBL Cinema SB231, 2.1 Channel Dolby Digital Soundbar with Wired Subwoofer for Deep Bass, Home Theatre with Remote, HDMI ARC, Bluetooth & Optical Connectivity (110W)

	A	B	C	D	E	F	G	H	I	
713	One speaker low	Sound missing	15 Apr 2021	Style name: Cinema	SB231 Colou	FALSE	VIKASH KUMAF	3	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
714	Aux jack availab	My tv has only h	19 Jan 2020	Style name: Cinema	SB110 Colou	FALSE	M A Khan	3	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
715	Waste money	It's not working i	20 Aug 2020	Style name: Cinema	SB261 Colou	FALSE	Deepak Indalkar	3	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
716	It is better to buy	The packing is n	14 Nov 2020	Style name: Cinema	SB110 Colou	FALSE	VASUDEVA SAF	2	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
717	Not 110watt	Not 110 Watt, so	16 Sep 2020	Style name: Cinema	SB110 Colou	FALSE	Santosh Gupta	2	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
718	Sounds not so g	There is no pow	13 Mar 2020	Style name: Cinema	SB110 Colou	FALSE	Tulsi Chugh	3	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
719	Not impressive	Very good sounc	04 Jan 2020	Style name: Cinema	SB110 Colou	FALSE	Krishna	3	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
720	Waste of money	Plz don't buy this	27 Jul 2020	Style name: Cinema	SB110 Colou	FALSE	sidagam srinivas	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
721	Buy when offer i	I feel just ok.. WI	21 Aug 2019	Style name: Cinema	SB110 Colou	FALSE	kura vinay	3	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
722	Not so good	Ok	08 Oct 2019	Style name: Cinema	SB110 Colou	FALSE	Srinadh Jyothi	2	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
723	Not good	30 watts Fake 11	12 Jul 2020	Style name: Cinema	SB110 Colou	FALSE	Amazon Custom	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
724	Jbl	Very bed sound	19 Oct 2019	Style name: Cinema	SB110 Colou	FALSE	Balkishan paray	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
725	Nice product	Nice product	17 Mar 2020	Style name: Cinema	SB110 Colou	FALSE	Siddhartha kum	3	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
726	Poor product	Poor sound the r	16 Aug 2019	Style name: Cinema	SB110 Colou	TRUE	Samrat Mukherji	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
727	Twice defective	Product receivec	25 Jun 2021	Style name: Cinema	SB110 Colou	TRUE	Vibhusita S.	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
728	Horrible JBL anc	Received a defa	07 Oct 2021	Style name: Cinema	SB110 Colou	TRUE	Suyash	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
729	waste of time,wa	im returning two	03 Mar 2020	Style name: Cinema	SB110 Colou	TRUE	Harsha	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
730	It better to buy lc	Speakers not wc	13 May 2020	Style name: Cinema	SB110 Colou	TRUE	DEVARASETTY	3	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
731	DO NOT BUY, W	It's been 2-3 mor	09 Mar 2020	Style name: Cinema	SB110 Colou	TRUE	Viresh G	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
732	Not worth it	Sound very muff	14 Jun 2020	Style name: Cinema	SB110 Colou	TRUE	Amazon Custom	2	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer

Fig. 4.2 Scrapped Data

- **Website** – amazon.com
- **Software for scrapping data** – python
- **Software used for keeping data in .CSV format** – Google Sheet

4.4 PROCESS OF FILTERING REVIEWS

Select Required Features For Sentiment Analysis

- **Features** – Title, content, date, variant, images, verified, author, rating, product, URL
- **Key Parameters**– title, content, verified, rating, product

	title	content	verified	rating	product
0	Why made in China ?	The product is good, build quality is premium,...	True	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
1	Worst customer care service by JBL	Worst customer care service. Placed a request ...	True	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
2	Worst customer care and worst product	Customer service is pathetic. The technical p...	True	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
3	A below average sound.	The sound is below average for a 200 watts sou...	True	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
4	best in bluetooth	sound via bluetooth is excellent but via hdmi ...	True	3.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...

Fig. 4.3 Imported Data

Drop Null Values

Here one merely deletes null values, or the records containing them, from the original data set. In case-wise deletion one deletes all records containing null values. In pair-wise deletion one only deletes records containing null values of variables used by a specific analysis.

Data Preprocessing (Tokenization, Lemmatization, Stemming, Stop Word Removal, Rejoining)

- **Tokenization –** Converting paragraphs into words.
- **Lemmatization –** Converting words into root words.
- **Stemming –** Reduced related words to common words.
- **Stop Word Removal –** Removing and, full stop, comma, etc.

❖ **Rejoining –** At last, rejoining the processed data.

	title	content	verified	rating	product
0	whi made in china	the product is good build qualiti is premium t...	1	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
1	worst custom care servic by jbl	worst custom care servic place a request for a...	1	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
2	worst custom care and worst product	custom servic is pathet the technic person is ...	1	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
3	a below averag sound	the sound is below averag for a watt sound bar...	1	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
4	best in bluetooth	sound via bluetooth is excel but via hdmi arc ...	1	3.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
...	...	...	...	...	...
718	horribl jbl and amazon	receiv a defaulti product sound issu difficult...	1	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
719	wast of timewast of money	im return two timeswhen i wa purcha but produc...	1	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
720	it better to buy local speaker	speaker not work properli left speaker work an...	1	3.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
721	do not buy wast of money a bit disappoint with...	it been month now and im use thi soundbar with...	1	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
722	not worth it	sound veri muffl	1	2.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...

723 rows × 5 columns

Fig. 4.4 Data after Preprocessing

4.5 DATA EXPLORATION

Presenting WORDCLOUD

- For review title

Fig. 4.5 WORDCLOUD for Review Title

- For review content

Fig. 4.6 WORDCLOUD for Review Content

Separating Dependent Variable and Target Variable

The target behavior which the intervention is designed to change. It depends on the environment to change it.

Here, dependent variable and target variable are separated in terms of :

- Ratings
- Product

Splitting Training And Testing Data

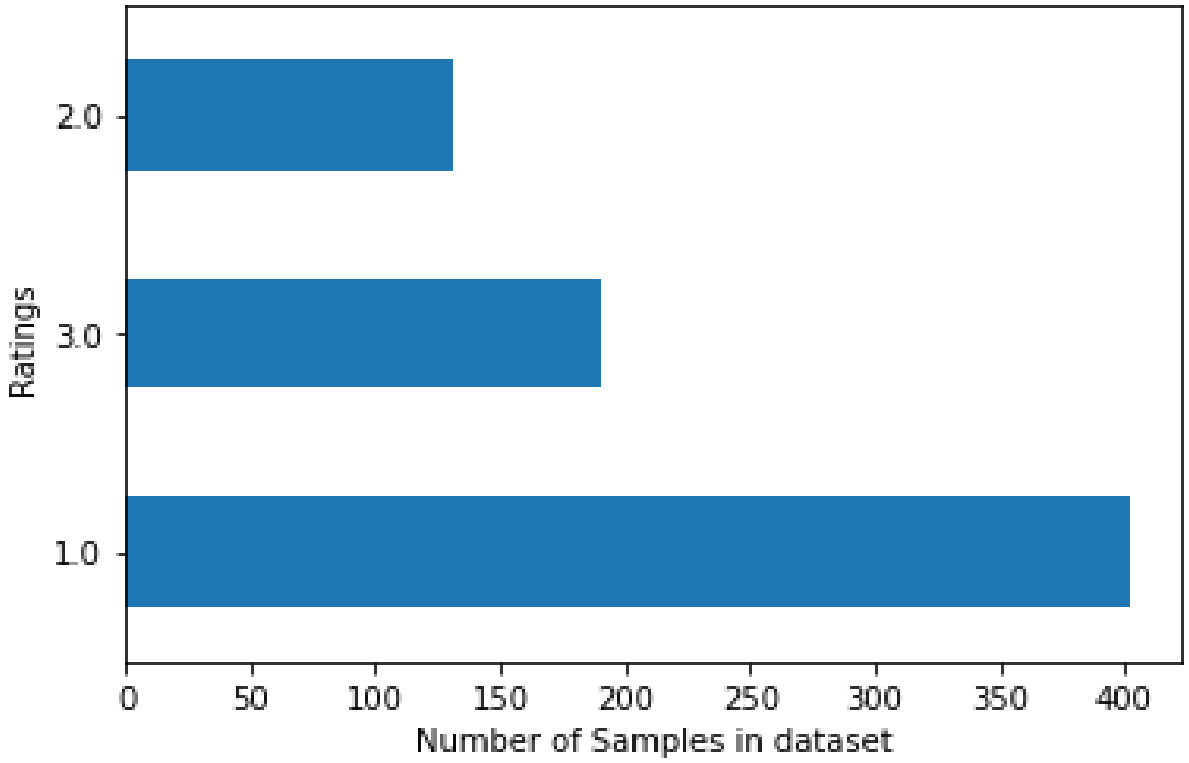


Fig. 4.7 Balance of Dataset

Splitting data into :

- X_train
- X_test
- Y_train
- Y_test

Checking Balance of Data

Checking balance of data in terms of :

- Ratings
- Number of samples in dataset.

4.6 CLASSIFICATION OF REVIEWS

Performing Fine Grained Sentiment Analysis

To understand about the sentiment present in the data entered by the customers, in accordance with their user experiences.

Applying Classification Algorithms

Applying classification algorithms such as :

- Logistic Regression
- Support Vector Machine
- Multinomial Naïve Bayes

Comparing Machine Learning Models

Comparing machine learning algorithms in terms of:

- Precision
- Recall
- F1-Score
- Support

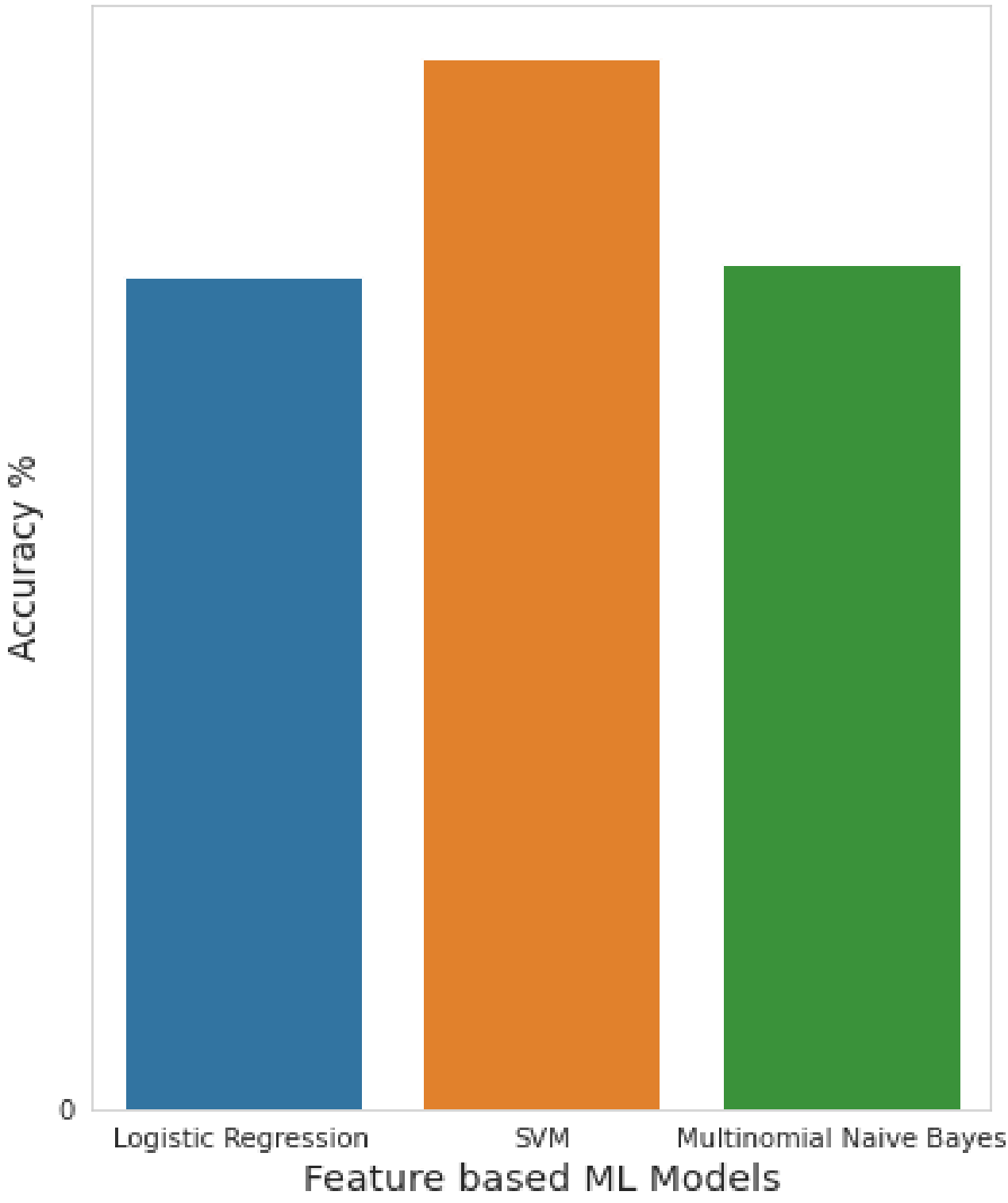


Fig. 4.8 Accuracy of Machine Learning Models

Performing Rule-Based Sentiment Analysis For Checking Polarity

Checking positive and negative polarity of title of reviews in terms of subjectivity_title and polarity_title.

	content	title	rating	subjectivity_title	polarity_title
0	the product is good build qualiti is premium t...	whi made in china	1.0	0.0	0.0
1	worst custom care servic place a request for a...	worst custom care servic by jbl	1.0	1.0	-1.0
2	custom servic is pathet the technic person is ...	worst custom care and worst product	1.0	1.0	-1.0
3	the sound is below averag for a watt sound bar...	a below averag sound	1.0	0.4	0.4
4	sound via bluetooth is excel but via hdmi arc ...	best in bluetooth	3.0	0.3	1.0

Fig. 4.9 Data for Polarity of Review

Plotting Histogram For Polarity Of Review Title

Presenting histogram in terms of :

- Polarity of review title
- Number of reviews (count)

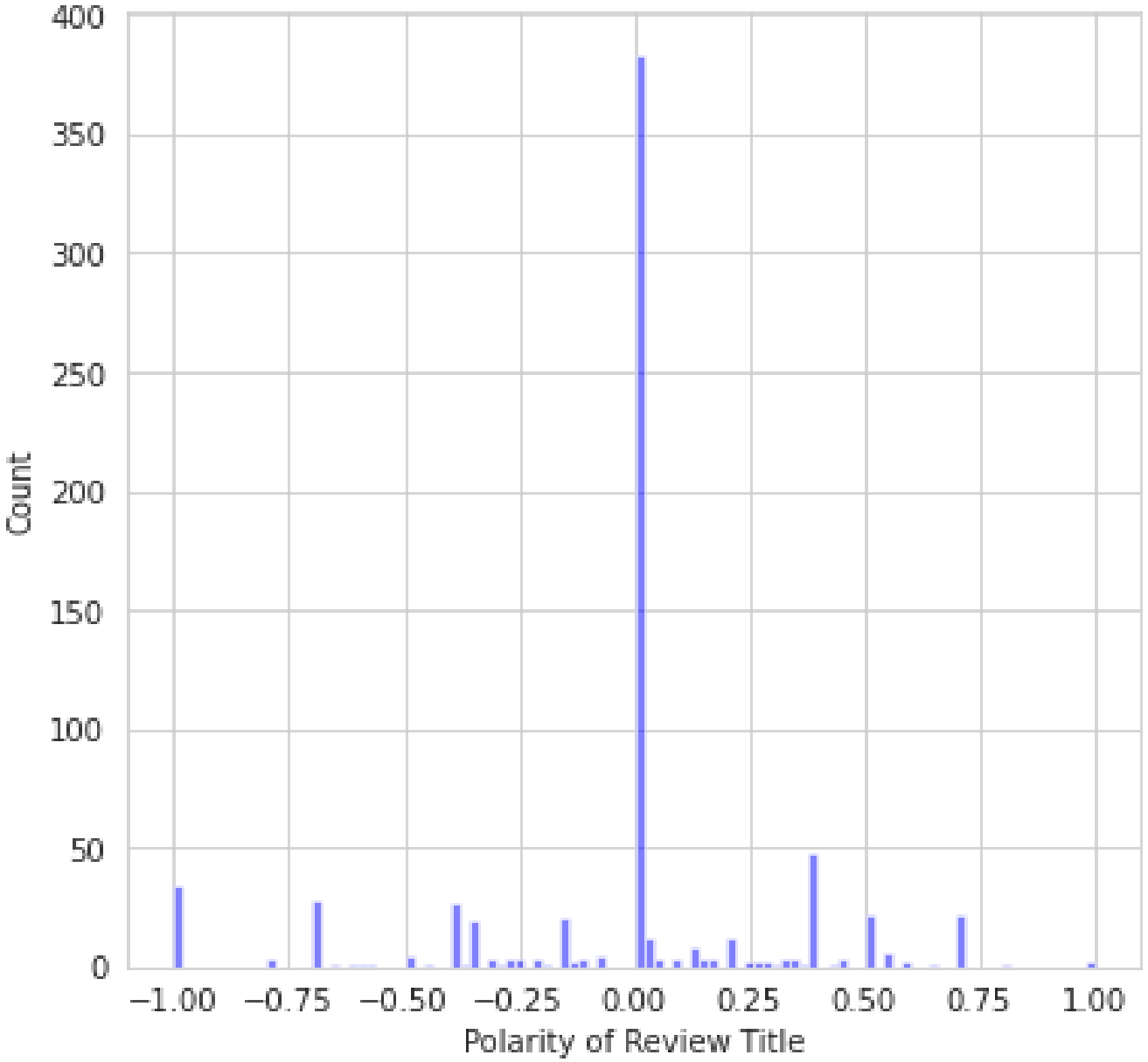


Fig. 4.10 Polarity of Review Title (Histogram)

Plotting histogram for polarity of review content

Presenting histogram in terms of :

- Polarity of review content
- Number of reviews (count)

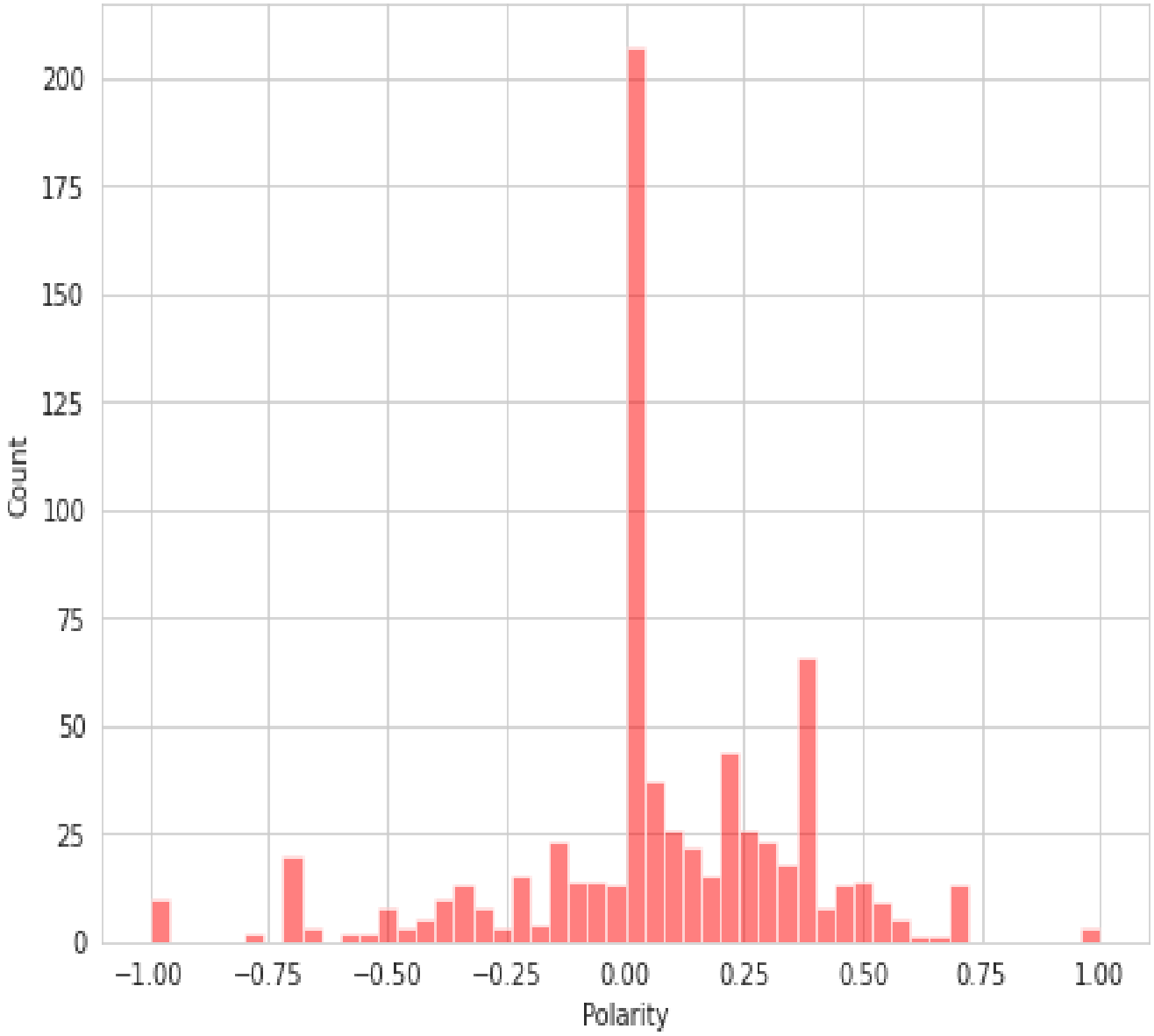


Fig. 4.11 Polarity of Review Content (Histogram)

Plotting box plot for polarity of title of review

Presenting box plot in terms of :

- Polarity of title
- Rating

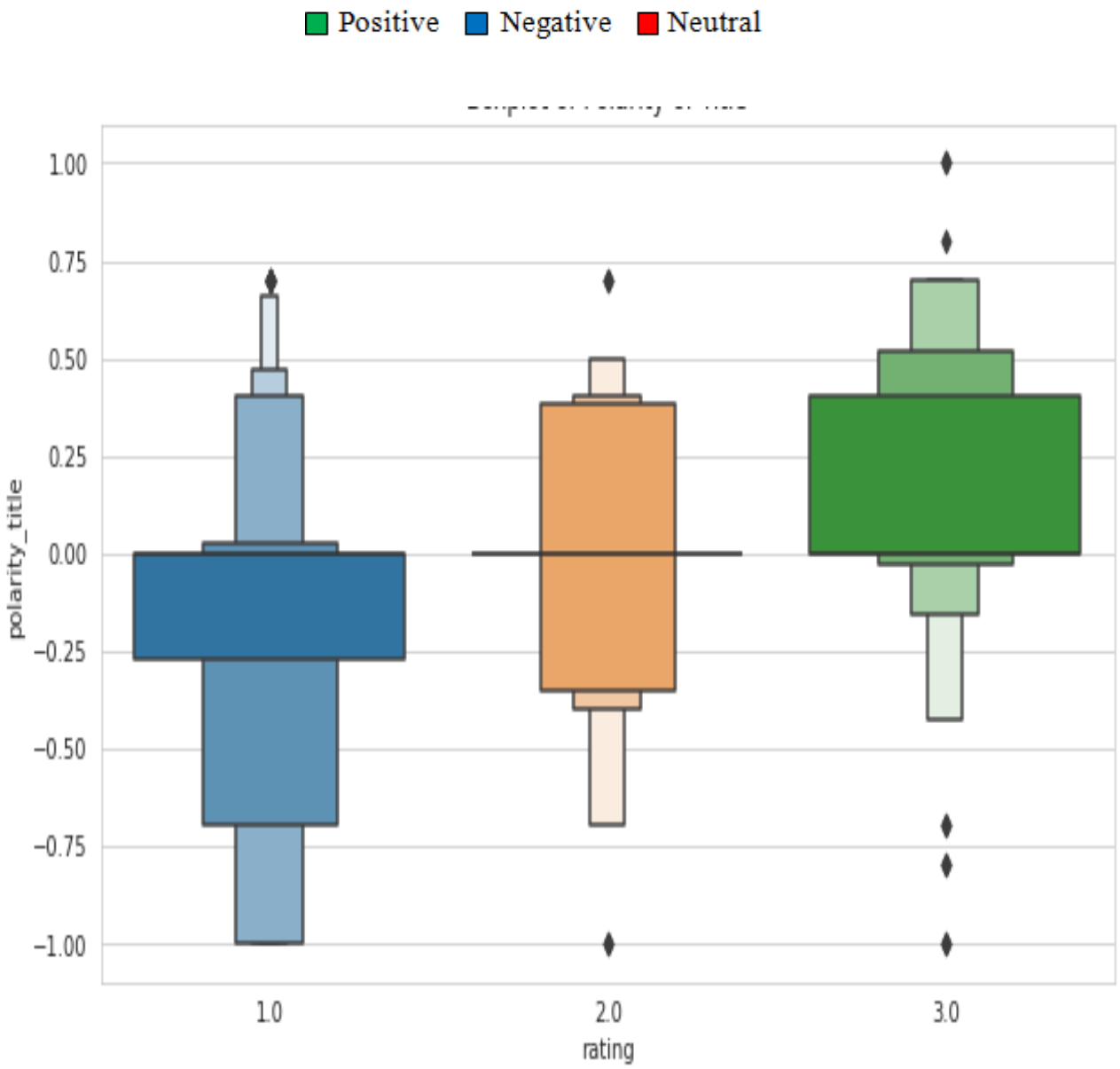


Fig. 4.12 Polarity of Review Title (Box Plot)

Plotting box plot for polarity of content of review

Presenting boxplot in terms of:

- Polarity of content
- Rating

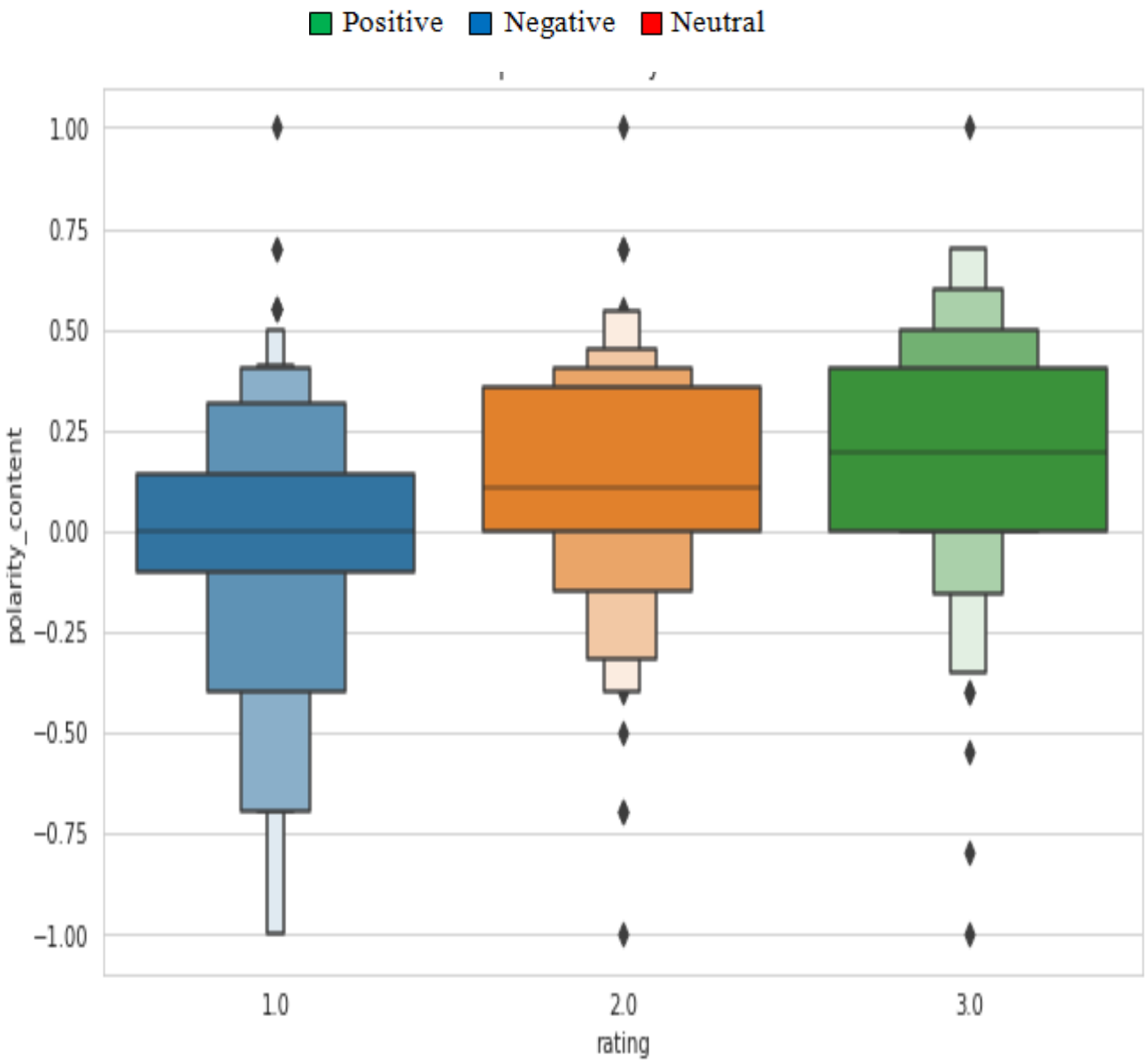


Fig. 4.13 Polarity of Review Content (Box Plot)

Plotting Density Plot And Histogram For Subjectivity Score Of Title Of Review

Presenting combined graph in terms of:

- Frequency
- Title Subjectivity

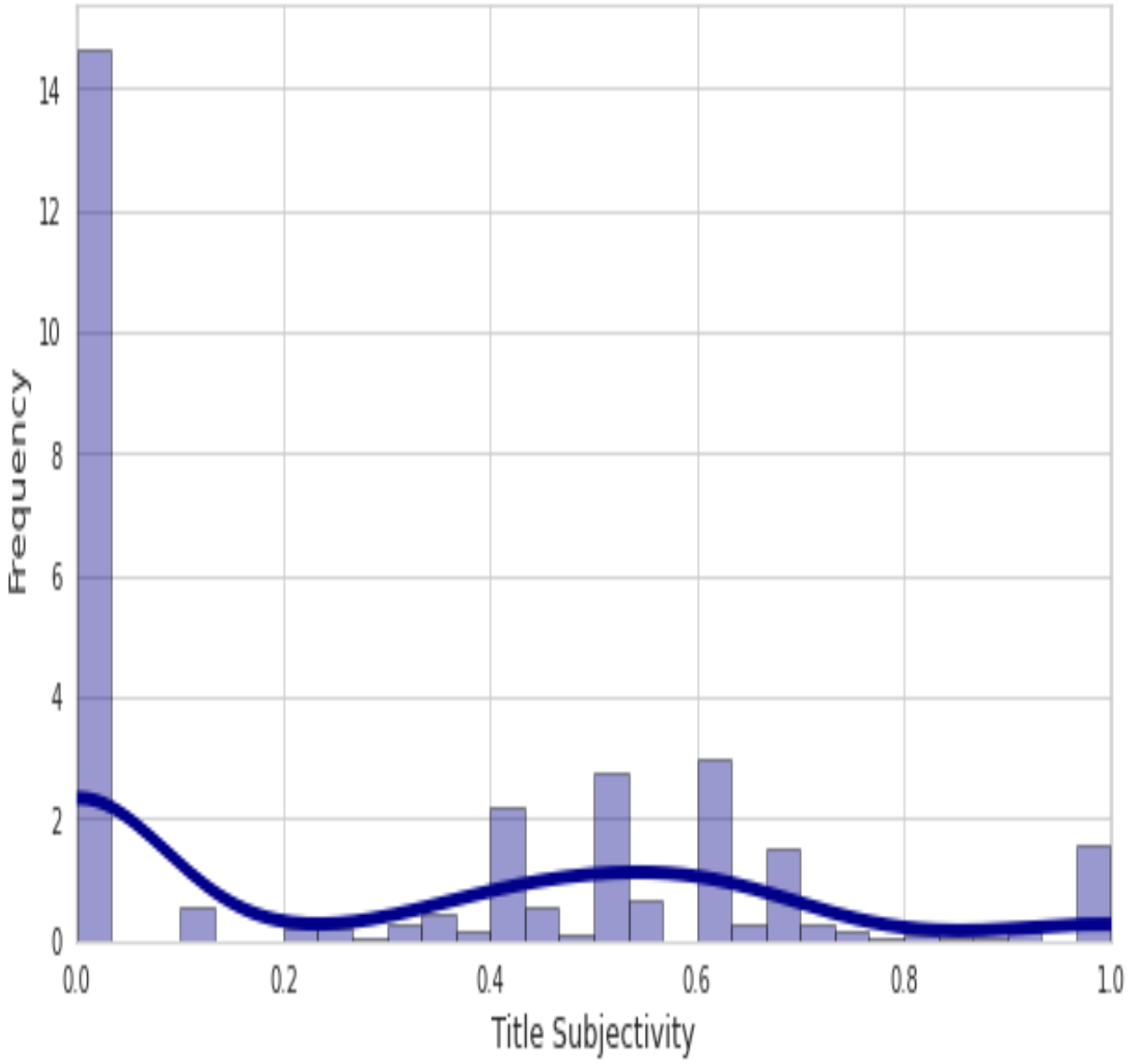


Fig. 4.14 Distribution of Title Subjectivity Score

Plotting density plot and histogram for subjectivity score of title of review

Presenting combined graph in terms of :

- Frequency
- Content Subjectivity

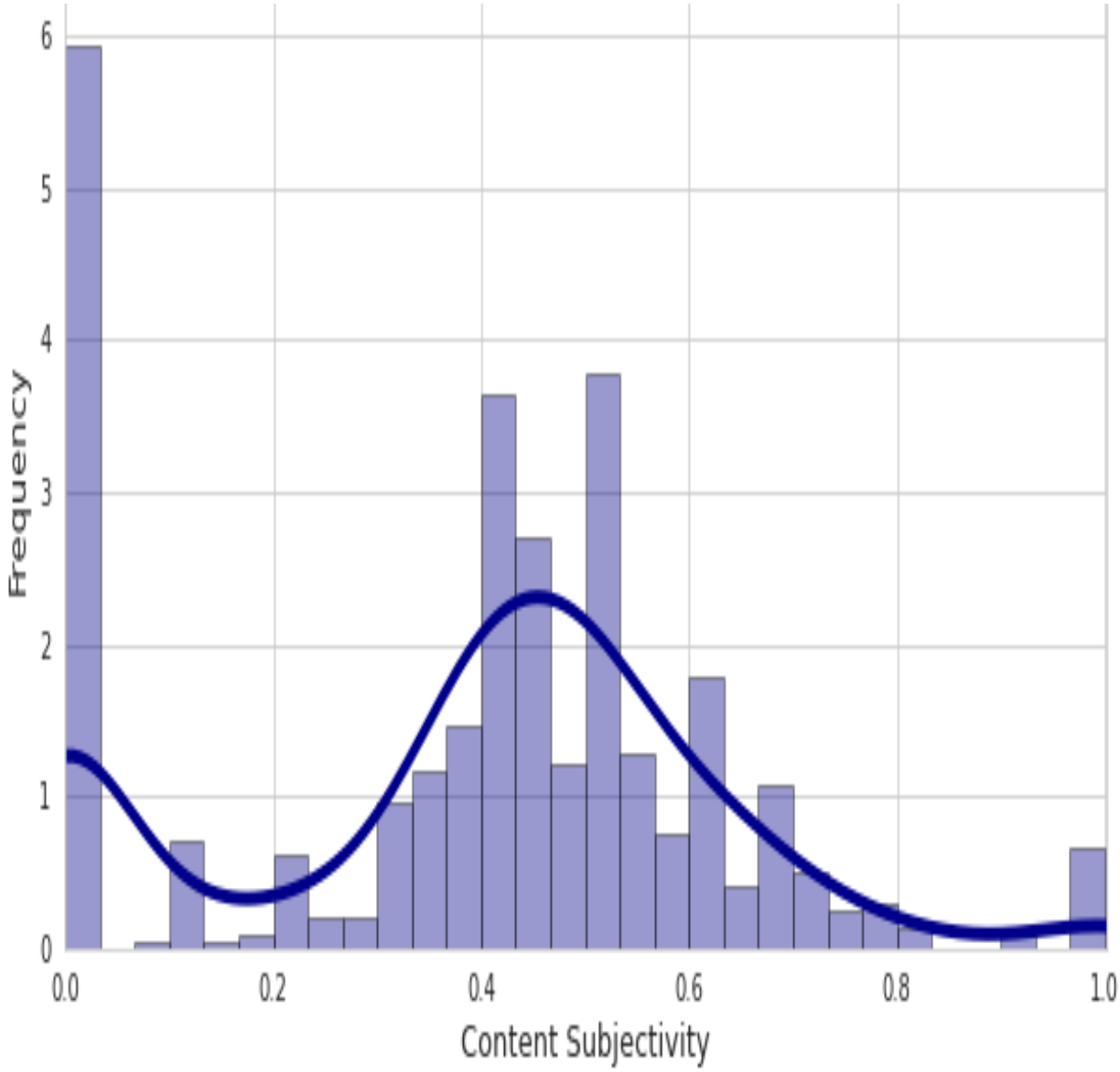


Fig. 4.15 Distribution of Content Subjectivity Score

4.7 DETECTION OF SENTIMENT

Performing VADER Analysis

After VADER analysis, the data is presented in terms of :

- Review Content
- Review title
- Subjectivity of title
- Polarity of title
- Polarity of content
- Negativity of title
- Negativity of content
- Positivity of title
- Positivity of content
- Subjectivity of content

	content	title	rating	subjectivity_title	polarity_title	subjectivity_content	polarity_content	neg_title	neg_content	neu_title	neu_content	pos_title
0	the product is good build qualiti is premium t...	whi made in china	1.0	0.0	0.0	0.446667	0.21250	0.000	0.031	1.000	0.831	0.000
1	worst custom care servic place a request for a...	worst custom care servic by jbl	1.0	1.0	-1.0	0.503571	-0.09881	0.363	0.027	0.354	0.815	0.283
2	custom servic is pathet the technic person is ...	worst custom care and worst product	1.0	1.0	-1.0	0.600000	-0.30000	0.569	0.083	0.208	0.917	0.222

Fig. 4.16 Data after VADER Analysis

CHAPTER 5

RESULT AND ANALYSIS

5.1 DATASET

The data was obtained from amazon.com.

5.1.1 Scrapped Data

	A	B	C	D	E	F	G	H	I
	title	content	date	variant	images	verified	author	rating	product	url
	Why made in Ch	The product is g	11 Aug 2020	Style name: Cine	https://images-na https://images-na https://images-na https://images-na https://images-na	TRUE	Gajanan	1	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
	Worst customer	Worst customer	23 Aug 2019	Style name: Cinema SB110 Colou		TRUE	Bhavesh Piparia	1	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
	Worst customer	Customer servic	02 Jan 2021	Style name: Cinema SB231 Colou		TRUE	Swarnalatha	1	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
	A below average	The sound is bel	10 Aug 2020	Style name: Cinema SB261 Colou		TRUE	Maharshi Magar	1	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
	best in bluetooth	sound via blueto	01 Aug 2019	Style name: Cinema SB110 Colou		TRUE	Krishna Prasad	3	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
	JBL's scam, sav	A totally wastefu	09 Sep 2020	Style name: Cine	https://images-na https://images-na	TRUE	mandeep kumar	1	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
	Very bad experie	Very bad produc	31 Dec 2019	Style name: Cinema SB110 Colou		TRUE	Anand ganesh p	1	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
	Not working with	Sound level very	22 Aug 2019	Style name: Cinema SB110 Colou		TRUE	nitin	2	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
	Third class chine	Its a third class c	22 Aug 2019	Style name: Cinema SB110 Colou		TRUE	Priyesh	1	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
	Remote malfunc	The jbl SB110 S	21 Aug 2019	Style name: Cine	https://images-na https://images-na	TRUE	ramei	1	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
	Disappointed wit	First of all It's no	03 Jun 2020	Style name: Cine	https://images-na	TRUE	Amazon Custom	1	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
	HDMI cable not	Sound is very g	16 Aug 2020	Style name: Cinema SB261 Colou		TRUE	AMIT	3	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
	ANNOYED WITH	The media could	28 Jun 2020	Style name: Cinema SB110 Colou		TRUE	manu j.	1	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
	Jbl soundbar sb	The device stopp	24 Aug 2019	Style name: Cinema SB110 Colou		TRUE	prithipal	2	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer

17 fx JBL Cinema SB231, 2.1 Channel Dolby Digital Soundbar with Wired Subwoofer for Deep Bass, Home Theatre with Remote, HDMI ARC, Bluetooth & Optical Connectivity (110W)

	A	B	C	D	E	F	G	H	I
713	One speaker low	Sound missing	15 Apr 2021	Style name: Cinema SB231 Colou		FALSE	VIKASH KUMAR	3	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
714	Aux jack availab	My tv has only h	19 Jan 2020	Style name: Cinema SB110 Colou		FALSE	M A Khan	3	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
715	Waste money	It's not working i	20 Aug 2020	Style name: Cinema SB261 Colou		FALSE	Deepak Indalkar	3	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
716	It is better to buy	The packing is n	14 Nov 2020	Style name: Cinema SB110 Colou		FALSE	VASUDEVA SAI	2	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
717	Not 110watt	Not 110 Watt, so	16 Sep 2020	Style name: Cinema SB110 Colou		FALSE	Santosh Gupta	2	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
718	Sounds not so g	There is no powe	13 Mar 2020	Style name: Cinema SB110 Colou		FALSE	Tulsi Chugh	3	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
719	Not impressive	Very good sounc	04 Jan 2020	Style name: Cinema SB110 Colou		FALSE	Krishna	3	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
720	Waste of money	Plz don't buy this	27 Jul 2020	Style name: Cinema SB110 Colou		FALSE	sidagam srinivas	1	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
721	Buy when offer i	I feel just ok.. WI	21 Aug 2019	Style name: Cinema SB110 Colou		FALSE	kura vinay	3	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
722	Not so good	Ok	08 Oct 2019	Style name: Cinema SB110 Colou		FALSE	Srinadh Jyothi	2	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
723	Not good	30 watts Fake 11	12 Jul 2020	Style name: Cinema SB110 Colou		FALSE	Amazon Custom	1	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
724	Jbl	Very bed sound	19 Oct 2019	Style name: Cinema SB110 Colou		FALSE	Balkishan paray	1	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
725	Nice product	Nice product	17 Mar 2020	Style name: Cinema SB110 Colou		FALSE	Siddhartha kuma	3	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
726	Poor product	Poor sound the r	16 Aug 2019	Style name: Cinema SB110 Colou		TRUE	Samrat Mukherji	1	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
727	Twice defective	Product receivec	25 Jun 2021	Style name: Cinema SB110 Colou		TRUE	Vibhusita S.	1	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
728	Horrible JBL anc	Received a defa	07 Oct 2021	Style name: Cinema SB110 Colou		TRUE	Suyash	1	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
729	waste of time,wa	im returning two	03 Mar 2020	Style name: Cinema SB110 Colou		TRUE	Harsha	1	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
730	It better to buy Ic	Speakers not wc	13 May 2020	Style name: Cinema SB110 Colou		TRUE	DEVARASETTY	3	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
731	DO NOT BUY, W	It's been 2-3 moi	09 Mar 2020	Style name: Cinema SB110 Colou		TRUE	Viresh G	1	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
732	Not worth it	Sound very muff	14 Jun 2020	Style name: Cinema SB110 Colou		TRUE	Amazon Custom	2	JBL Cinema SB2	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer

Fig. 5.1 Raw Data

5.1.2 CLEANED DATA

	title	content	verified	rating	product
0	whi made in china	the product is good build qualiti is premium t...	1	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
1	worst custom care servic by jbl	worst custom care servic place a request for a...	1	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
2	worst custom care and worst product	custom servic is pathet the technic person is ...	1	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
3	a below averag sound	the sound is below averag for a watt sound bar...	1	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
4	best in bluetooth	sound via bluetooth is excel but via hdmi arc ...	1	3.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
...	...	...	...	...	...
718	horribl jbl and amazon	receiv a defaulti product sound issu difficult...	1	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
719	wast of timewast of money	im return two timeswhen i wa purcha but produc...	1	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
720	it better to buy local speaker	speaker not work properli left speaker work an...	1	3.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
721	do not buy wast of money a bit disappoint with...	it been month now and im use thi soundbar with...	1	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
722	not worth it	sound veri muffl	1	2.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...

723 rows × 5 columns

Fig. 5.2 Cleaned Data

5.2 WORDCLOUD GENERATED

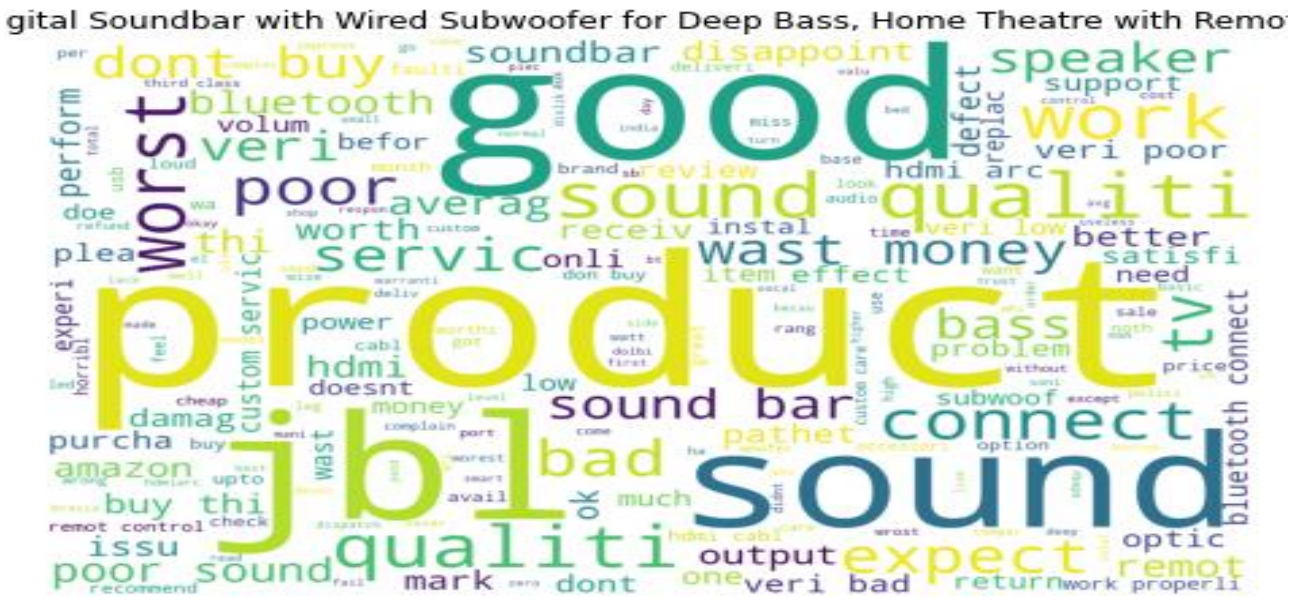


Fig. 5.3 Word Cloud

5.3 ACCURACY EVALUATION OF ML MODELS

S No.	ML Model	Train Accuracy	Test Accuracy
1	Naïve Bayes	89.43%	89.91%
2	SVM	95.56%	94.91%
3	Logistic Regression	87.13%	88.79%

Table 5.1 Accuracy Evaluation

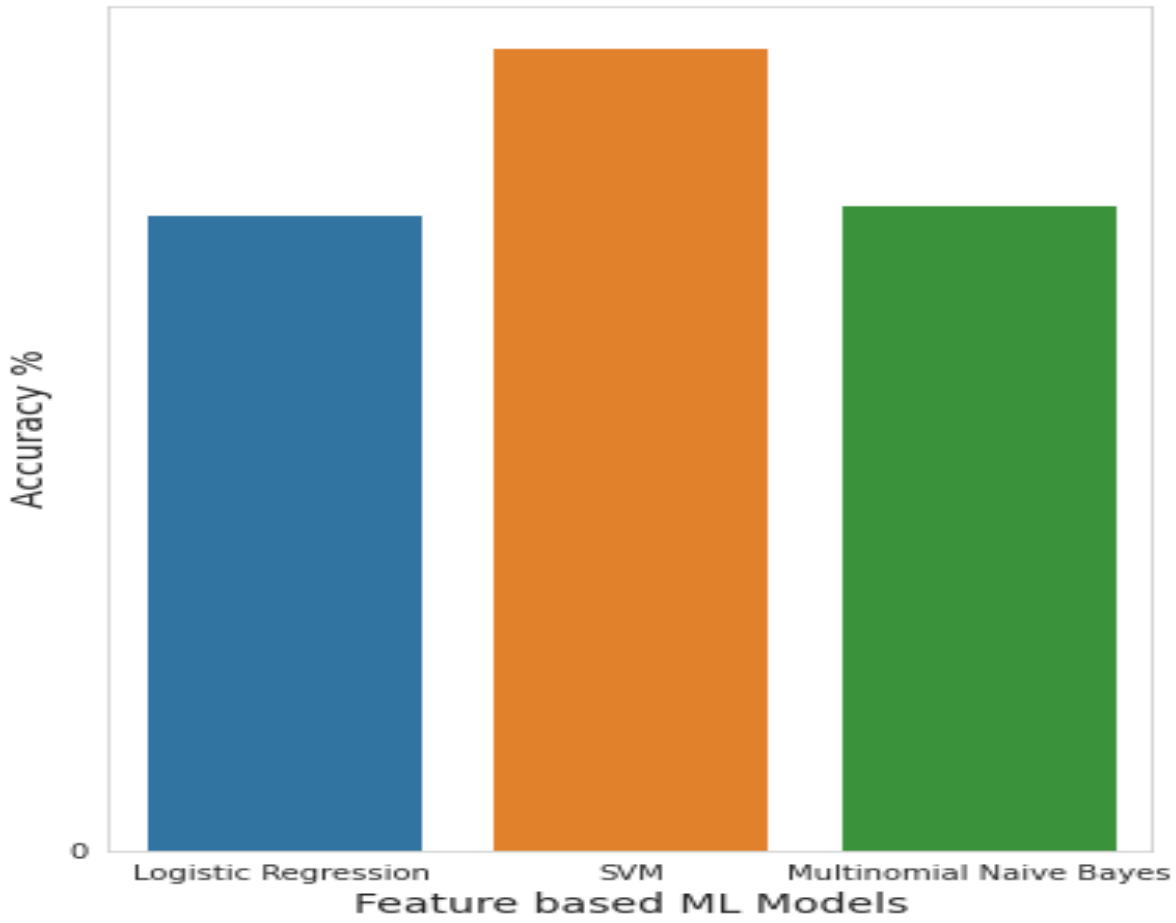


Fig. 5.4 Accuracy of Machine Learning Models

5.4 POLARITY OF REVIEW

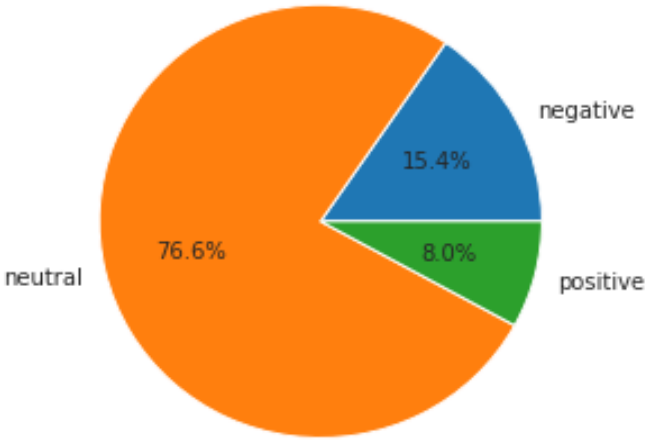


Fig. 5.5 Polarity (Pie Chart)

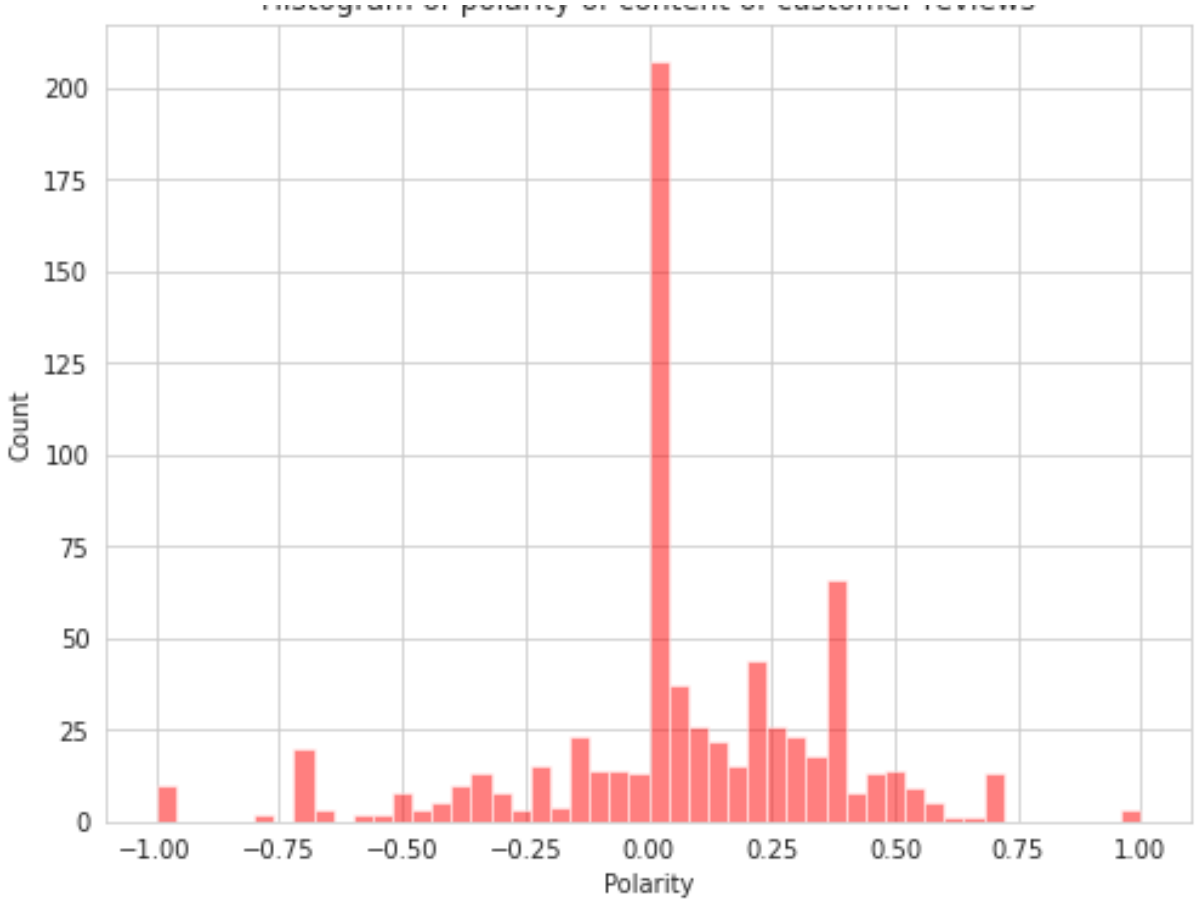


Fig. 5.6 Polarity (Histogram)

5.5 PRODUCT RECOMMENDATION

According to the result generated, the majority of the user had shown the neutral behavior towards the product, therefore, it is recommended that the product is **risky** to buy.

5.6 ANALYSIS AND DICUSSION ON RESULT

Amazon is the world's largest e-commerce platform and hence, it is a pool of reactions, sentiments of customers as well. Thus, we can observe their feelings, sentiments, emotions towards the product from this site. Like this, user can opt for any product from this site to get a hand on its reviews. User just needs to copy the URL of the product through the website.

So, not only the product used in this work, we can retrieve sentiments of any product just by using product's URL .

CHAPTER-6

CONCLUSION AND FUTURE SCOPE

6.1 CONCLUSION

With many applications, sentiment analysis is a rapidly growing field. Not only can consumer expectations be fulfilled based on the outcome of the sentiment analysis, but suppliers, distributors, etc. can also get an idea of the user or client's reaction and therefore ensure that they can make and meet the required adjustments. Sentiment analysis has been an important tool for brands looking to learn more about how their customers are thinking and feeling. We have studied different methods and approaches of ML. The techniques of machine learning are much simple and easy to incorporate. These approaches achieve critical outcomes.

The system proposed in this report aims to help a user select a JBL Sound Bar based on his/her needs using review data from previous users. It pulls review data from the Amazon website periodically, processes it and assigns a score to each feature of each phone based on the review data. When the user inputs his/her preferences, the scores are used to determine the best match for the user. This match is guaranteed to be up-to-date.

A model was tested by using Support Vector Machine, Naïve Bayes and Logistic Regression on datasets of product reviews to find the polarity of sentiments and texts whether positive, negative or neutral. The performance resulting models tested to obtain the value of Accuracy, Recall, Precision, and F-1 measure of all three models used. Finally the Support Vector Machine (SVM) algorithm has been achieved higher accuracy, i.e., 90.99% and it is found that the SVM is a robust and better one.

In summary, proposed work tried Naive Bayes, SVM and Logistic Regression. Research is supposed to provide more flexible and accurate solution. The proposed research is supposed to resolve the issue of previous research that was faced during sentiment analysis.

6.2 FUTURE SCOPE

Amazon is the world's largest e-commerce platform and hence, it is a pool of reactions, sentiments of customers as well. Thus, we can observe their feelings, sentiments, emotions towards the product from this site. Like this, user can opt for any product from this site to get a hand on it's reviews. User just need to copy the URL of the product through the website. So, not only the product used in this work, we can retrieve sentiments of any product just by using product's URL .

Sentiment analysis is a uniquely powerful tool for businesses that are looking to measure attitudes, feelings and emotions regarding their brand. To date, the majority of sentiment analysis projects have been conducted almost exclusively by companies and brands through the use of social media data, survey responses and other hubs of user-generated content. By investigating and analyzing customer sentiments, these brands are able to get an inside look at consumer behaviors and, ultimately, better serve their audiences with the products, services and experiences they offer.

The future of sentiment analysis is going to continue to dig deeper, far past the surface of the number of likes, comments and shares, and aim to reach, and truly understand, the significance of social media interactions and what they tell us about the consumers behind the screens. This forecast also predicts broader applications for sentiment analysis – brands

will continue to leverage this tool, but so will individuals in the public eye, governments, nonprofits, education centers and many other organizations.

Further, there are many more areas where future research and development can excel, like-

- **Deeper, Broader Insights from Sentiment Analysis**

Sentiment analysis is getting better because social media is increasingly more emotive and expressive. A short while ago, Facebook introduced "Reactions," which allows its users to not just 'Like' content, but attach an emoticon, whether it be a heart, a shocked face, angry face, etc. To the average social media user, this is a fun, seemingly silly feature that gives him or her a little more freedom with their responses. But, to anyone looking to leverage social media data for sentiment analysis, this provides an entirely new layer of data that wasn't available before. Every time the major social media platforms update themselves and add more features, the data behind those interactions gets broader and deeper.

- **Greater Personalization for Audiences**

As a result of deeper and better understanding of the feelings, emotions and sentiments of a brand or organization's key, high-value audiences, members of these audiences will increasingly receive experiences and messages that are personalized and directly related to their wants and needs. Rather than segment markets based on age, gender, income and other surface demographics, organizations can further segment based on how their audience members actually *feel* about the brand or how they use social media. While some people shudder at the thought of companies learning more about them, more exact targeting means that, in the near future, we will no longer be scratching our head wondering why we see advertisements for products we'd never dream of purchasing. In

other words, the spray-and-pray advertising tactics are almost put to rest and there will be a time when every marketing message we see will be relevant and useful to us. Sentiment analysis is going to be a large contributing factor towards achieving this vision.

- **Not Just For Marketers and Brands**

Again, sentiment analysis is on the verge of breaking into new areas of application. While we will likely always think of it first in terms of the traditional marketing sense, the world has already seen a few ways that sentiment analysis can be used in other areas. Social media analytics helped predict and explain the emotions of concerned parties behind Brexit and the 2016 US election, which has spurred a number of non-brand organizations to investigate how sentiment analysis can be used to predict outcomes and map out the emotional landscape of people, voters and the like. Additionally, businesses are looking at ways that sentiment analysis can be used outside of their marketing and PR departments. Sentiment analysis simply looks more popular in the future.

- **Algorithm-Based Sentiment Analysis Plateaus**

Algorithms have long been at the foundation of most forms of analytics, including social media and sentiment analysis. With recent years bringing big leaps in machine learning and artificial intelligence, many analytics solutions are looking to these technologies to replace algorithms. Unfortunately for organizations looking to leverage sentiment analysis to measure audience emotions, machine learning isn't yet ready to tackle the complex nuances of text and how we talk, especially on social media channels that are rife with slang, sarcasm, double meanings and misspellings. These make it difficult for artificial intelligence systems to accurately sort and classify sentiments on social media. And, with

any analysis project, accuracy is crucial. It is uncertain if machine learning will progress to the point that it *is* capable of accurately analyzing text, or if sentiment analysis projects will have to find a new basis to avoid the current plateau of algorithms. Some social media analytics solutions have begun taking a more human approach to deciphering the often ambiguous nature of text, but this can be time consumin

REFERENCES

[1] Raj Sinha, “Data analysis and sentiment analysis on Amazon reviews”, International Journal for Research in Applied Science and Engineering Technology [IJRASET], volume-9, pp. 2200-2206, 2021.

[2] Arwa S. M. AlQahtani, “Product sentiment analysis for Amazon reviews”, International Journal of Computer Science and Information Technology [IJCSIT], volume 13, pp.15-30, 2021.

[3] Somsurva Dutta and Santosh Bothe, “Analysis of Amazon reviews using machine learning approach”, International Journal for Research in Applied Science and Engineering Technology [IJRASET], volume-9, pp. 313-323, 2021.

[4] Xing Fang and Justin Zhan, “Sentiment analysis using product review data”, Journal of Big Data [JBD], 2015.

[5] Waqar Muhammad, Khurum Nazir Junejo, Maria Mushtaq and Muhammad Yaseen Khan, “Sentiment analysis of product reviews in the absence of labeled data using supervised learning approaches”, Research Gate, 2019.

[6] Najma Sultana, Sourabh Chandra, Pintu Kumar and Sk Safikul Alam, “Sentiment analysis for product review”, Research Gate, 2019.

[7] Pravesh Kumar Singh, “Analytical study of feature extraction techniques in opinion mining”, Research Gate, pp.85-94, 2013.

[8] Minu P Abraham and Udaya Kumar Reddy, “Feature based sentiment analysis of mobile product reviews using machine learning techniques”, International Journal of

Advanced Trends in Computer Science and Engineering [IJATCSE], volume-9, pp. 2289-2296, 2020.

[9] Tanjim Ul Haque, Nudrat Nawal Saber and Faisal Muhammad Shah, “Sentiment analysis on large scale Amazon product reviews”, IEEE-International Conference of Innovative Research and Development [ICIRD], 2018.

[10] Raheesa Safrin, K.R.Sharmila and T.S.Shri subangi, “Sentiment analysis on online product review”, International Research Journal of Engineering and Technology [IRJET], volume -4, pp. 2381-2388, 2017.

[11] Rajkumar S Jagdale, Vishal S Shrisat and Sachin N Deshmukh, “Sentiment analysis on product reviews using machine learning techniques”, Springer, pp 639-647, 2018.

[12] T.K Shivaprasad and Jyothi Shetty, “Sentiment analysis of product reviews”, IEEE, 2017.

[13] Prashant Pandey, Muskan and Nitasha Soni, “Sentiment analysis on customer feedback data”, IEEE, 2019.

[14] Monir Yahya Ali Salmony and Arman Rasool Faridi, “Supervised sentiment analysis on Amazon product reviews”, IEEE, 2021.

[15] Anjana Madhav C and Lavanya M, “Sentiment analysis of product reviews for overall product rating”, IEEE, 2020.

[16] Duvvuru Mahammad Dawood Khan, “Sentiment analysis of product based reviews”, [IJIRT], volume-8, pp. 467-473, 2021.

[17] Panthati Jagadeesh, Ranga Tarun Kumar, Challa Manish Reddy and Jasmine T. Bhaskar, “Sentiment analysis of product reviews”, Research Gate, 2017.

[18] P Rakesh, M Sandeep and G Jagadeesh, “Amazon product review sentiment analysis using machine learning”, International Research Journal of Computer Science [IRJCS], volume-8, pp.136-141, 2021.

[19] Arpita Lasod and Rahul Pawar, “Sentiment analysis using machine learning techniques”, International Journal of Innovative Research in Technology [IJIRT], volume-6, pp.153-157, 2019.

[20] K Ashok Kumar, “Sentiment analysis of Amazon product reviews using machine learning”, Research Gate, volume-82, pp.5245-5254, 2020.

[21] Kiran Shehzadi and Usman Ahmed Raza, ”Sentiment analysis by using deep learning and machine learning techniques”, International journal of Advanced Trends in Computer Science and Engineering [IJATCSE], volume-10, pp.754-761, 2021.

[22] Sobia Wassan, Xi Chen, Tian Chen, Muhammad Waqr and N Z Jhanjhi, ”Amazon product sentiment analysis using machine learning techniques”, Research Gate, volume 30, pp.695-703, 2021.

[23] Vineet Jain and Mayuri Kambli, ”Amazon product reviews: Sentiment analysis”, Research gate, 2020.

[24] Jyoti Budhwar,“Sentiment analysis based method for Amazon product reviews”, International Journal of Engineering Research and Technology [IJERT], volume-9, pp.54-57, 2021.

[25] Sayyed Johar and Samara Mubeen, “Sentiment analysis on large scale Amazon product reviews”, International Journal of Science Research in Computer Science and Engineering [IJSRCSE], volume-8, pp.07-15, 2020.

[26] P. Rakesh, M. Sandeep and G. Jagadesh, “Amazon product review sentiment analysis using machine learning”, International Research Journal of Computer Science [IRJCS], volume-08, pp.136-141, 2021.

[27] I Kaur and G Lal, “Sentiment analysis of Amazon canon camera review using hybrid model”, International Journal of Computer Application [IJCA], volume-182, pp.25-32, 2018.

[28] Ion Smeureanu and Cristian Bucur, “Applying supervised opinion mining techniques on online user reviews”, Informatica Economica, Volume-16, 2012.

[29] Bo Pang and Lillian Lee, “Opinion mining and sentiment analysis”, Foundations and Trends_R in Information Retrieval, volume-2, 2015.

PUBLICATIONS

PUBLICATIONS FROM THIS WORK

1) **"Evaluation and Visual Representation of Online Products Using Machine Learning based Sentiment Analysis"** has been published in International journal Of Creative Research Thoughts, volume-10, Issue-7 of July 2022.

(http://ijcrt.org/viewfull.php?&p_id=IJCRT2207275)

CERTIFICATE OF PUBLICATION

IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

VISUAL REPRESENTATION OF PRODUCT REVIEWS USING MACHINE LEARNING AND SENTIMENT ANALYSIS

[1]Surabhi Agarwal , [2]Mohd Usman Khan

[1]P.G Student, CSE Department, Integral University, Lucknow, U.P, India

[2]Assistant Professor, CSE Department, Integral University, Lucknow, U.P, India

***Abstract*: Presently, very huge amount of data is available on internet. This data holds expressed opinions and sentiments. The volume, variety and velocity are the key properties of this data. Decision making on both individual and organizational level is always accompanied by the search of other's opinions on the same. With the tremendous establishment of opinion rich resources like product reviews, feedbacks are proved to be the most essential and valuable resources to market. Sentiment Analysis is an application of Natural Language Processing (NLP), also known as emotion extraction or opinion mining or text mining. It helps to understand the human decision making, categorizing, analyzing and extracting meaningful information in order to understand opinions of consumers. There are several tools and algorithms available to perform sentiment detection and analysis, which are better than unconventional, time consuming and error prone methods used earlier.**

Index Terms - Sentiment Analysis, Opinion Mining, Text Analysis, Natural Language Processing (NLP), Product Review, Data Classification, Polarity Detection.

I. INTRODUCTION

The advancement of electronic commerce with growth in internet and network technologies has led customers to move to online retail platforms such as Amazon, Walmart, etc. People usually rely on customer reviews of products before they buy online. These reviews are often rich in information describing the products and their quality. Customers choose to compare between various products and brands based on whether an item has a positive or negative review. These reviews act as a feedback mechanism for the seller. Through this medium, sellers strategize their future sales and the areas where the product or services needs improvement.

The enormous amount of competition to attract and maintain customers online is fascinating businesses to implement novel strategies to enhance the customer experiences. It is becoming compulsory for companies to examine customer reviews on online platforms such as Amazon to understand better how customers rate their products and services. The purpose of this study is to investigate how companies can conduct sentiment analysis based on Amazon reviews to gain more intuitions into customer experiences. The dataset selected for this research consists of customer reviews of Amazon products, which enables a business person to gain insights on customer reviews regarding specific product and services. The study will enable companies to pinpoint the reasons for positive and negative reviews, followed by implementing effective strategies to address them accordingly. The aim of this research is to help companies to use sentiment analysis to understand customer experiences and customers to understand whether a particular product is to be purchased or not.

II. LITERATURE SURVEY

Sentiment analysis has been present for some time, but many active researches had happened in the past few years to understand and exhibit customer reviews.

Levent Guner [1] from KTH Royal Institute of Technology, Stockholm selected 60,000 random product reviews from Amazon. He used the dataset available in Kaggle that contains 4 million reviews. The performance was compared with three different algorithms namely Naïve Bayes (NB), Support Vector Machine (SVM) and Long short-term memory network (LSTM). The authors used numerous performance metrics to determine the best performing classification algorithm on the test set. To determine the performance, the metrics used were Accuracy, Area Under Curve (AUC), Precision, Recall and F1-score. Based on the results of the evaluation, their study concluded that the LSTM model performed the best with precision > 0.90 and AUC = 0.96 for binary classification.

Xing Fang and **Justin Zhan** [2] collected over 5.1 million product reviews in 4 key categories: beauty, book, electronics, and home. They analyzed these reviews with 3 different classifiers, namely, Naïve Bayes, Support Vector Machine and Random Forest. Their paper addressed the basic question of judging sentiments, categorizing sentiment polarity and ends with random forest generating more accurate results. As per their findings, for larger data sets SVM worked better than Naïve Bayes.

Wan Liang Tan [3] performed both traditional machine learning algorithms including Naïve Bayes, SVM, K-Nearest Neighbor and Deep Learning Network Models such as Recurrent Network Models and LSTM on Amazon reviews dataset. They collected 34627 reviews and divided it into 21000 records of training datasets and 13627 test datasets respectively. In terms of test accuracy, LSTM showed the best performance among all of them with 71.5 percent accuracy. One of the main reasons for not high enough accuracy was the imbalance in their data, as they concluded in their work.

Callen Rain [4] used Naïve Bayes and Decision-List classifiers to sort product reviews from Amazon as positive and negative. He used a corpus that includes 50,000 reviews of 15 items that is used as the research dataset. The features such as bag-of-words and bigrams are compared with each other for labeling positive and negative reviews correctly. His analysis showed that Naive Bayes performed better than the decision-list and bag of words ended up being the best algorithm for feature extraction.

Nishit Shrestha and **Fatma Nasoz** [5] analyzed the reviews present on Amazon to get opinions. They had a model using Recurrent Neural Networks (RNN) with Gated Recurrent Unit (GRU) that learned low dimensional review vector representation using paragraph vectors and product embedding. The data used in this analysis is a collection of about 3.5 million product reviews gathered from Amazon.com. Paragraph Vectors are very much inspired by word vectors. PV system learns vectors by estimating the next term, given several sampled contexts from a paragraph. The concatenation of review embedding developed from paragraph vectors and GRU-derived product embedding is used to train a Support Vector Machine (SVM) to exhibit sentiments. With only review embedding, the anticipated classifier provided 81.29 percent accuracy. The product embedding techniques improved the accuracy to 81.82 percent. Authors assume that a similar technique can be used to acquire user information.

In a **research article** [6] different approach has been implemented for sentimental analysis. In this research, an algorithm called a BoW (Bag of words) is used in which the relation between the words is not taken in account. To measure the sentiment for the whole sentence, the sentiment of every single word of the sentences has been individually decided and values are collected using some aggregation of function. Along with this opinion summarization method based on features can also be used. For each product, a specific feature and their attributes are extracted, and the general feature for each product class is acquired. Then polarity is assigned to each function with the aid of Sequential Minimal Optimization and Support Vector Machines.

III. OBJECTIVE

- To provide visual representation of reviews in the form of word cloud, histogram, box plot.
- To provide visual representation of sentiments of reviews in the form of pie-chart
- To provide product recommendation to users.

IV. SYSTEM DISTRIPTION MODEL

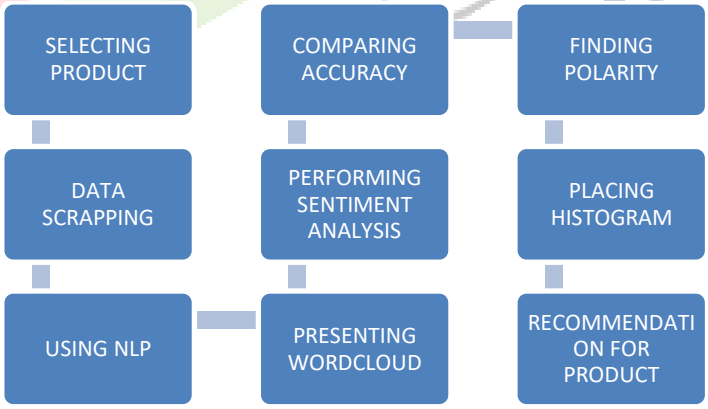


Fig. Model for Sentiment Analysis

4.1 NATURAL LANGUAGE PROCESSING

NLP is a branch of AI that helps computers to understand, interpret and manipulate human languages like English or Hindi to analyze and derive it's meaning. NLP helps developers to organize and structure knowledge to perform tasks like translation, summarization, named entity recognition, relationship extraction, speech recognition, topic segmentation, etc.
NLP aims at converting unstructured data into computer-readable language by following attributes of natural language. Machines employ complex algorithms to break down any text content to extract meaningful information from it. The collected data is then used to further teach machines the logics of natural language. Natural language processing uses syntactic and semantic analysis to guide machines by identifying and recognizing data patterns.

The natural Language Processing procedure is as follows –

4.1.1 Data Collection- The very first job in the process of sentiment analysis is data collection. Data can be collected from various sources like any website, from the several online opinion sets & ratings.

4.1.2 Data Preprocessing- It is the cleaning process of data. Unrequired words & symbols are omitted. This is required for further processing to be streamlined. Part of this move is eliminating hyperlinks, repeated sentences, emoticons, and special characters. It also performs lemmatization and stemming. Finally, it takes a reduced collection of features and passes them to the classifiers.

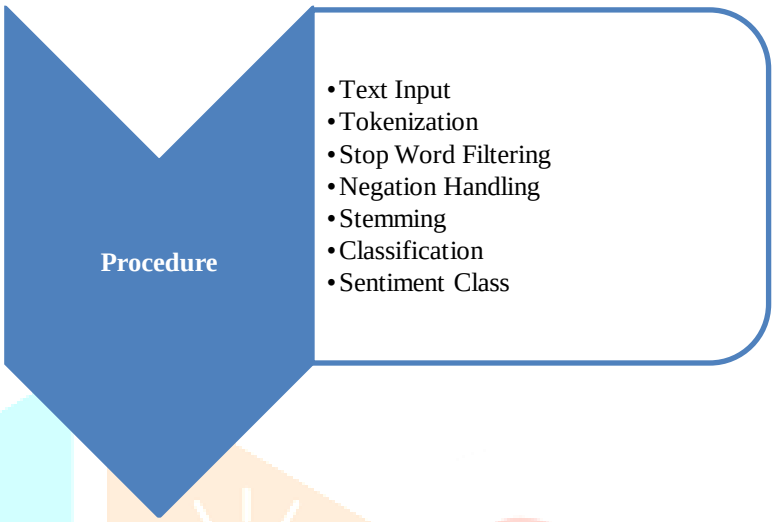


Fig. Natural Language Processing Procedure

4.1.3 Classification- The most critical aspect of a system for sentiment analysis is a classifier. Classification is achieved in negatives, positive, or neutrals categories. A third of the database is usually used as training sets to generate the classifiers. To a large degree, the precision of the classifier relies on the training collection. By using machine learning classifiers like SVM, Bayesian Classifiers and so on, the classification can be performed. However, before training and testing the classifier, machine learning classifiers do feature extraction, which can also use deep neural networks for classifying the data.

4.1.4 Output- After the data has gone through the classifier, the output data is shown. It shows the polarity of feelings of the whole data, and the degree of detail depends upon the type of classifiers which is used. The output can be represented in the form of a word cloud, histogram, box plot, pie chart, graph. They help user to understand reviews easily without spending too much time in scrolling down the list of reviews.

4.2 WORD CLOUD

Word clouds (also known as text clouds or tag clouds) work in a very simple way that the more a specific word appears in a source of text data (such as a speech, blog, post, or database), the bigger and bolder it appears in the word cloud. A word cloud is a collection, or cluster, of words represented in different sizes. The bigger and bolder the word appears, the more often it is raised up within a given text and the more important it is. Also known as tag clouds or text clouds, these are ideal ways to pull out the most relevant parts of textual data, from blog posts to databases. They can also help business users to compare and contrast two different pieces of text to find the word related similarities between the two.

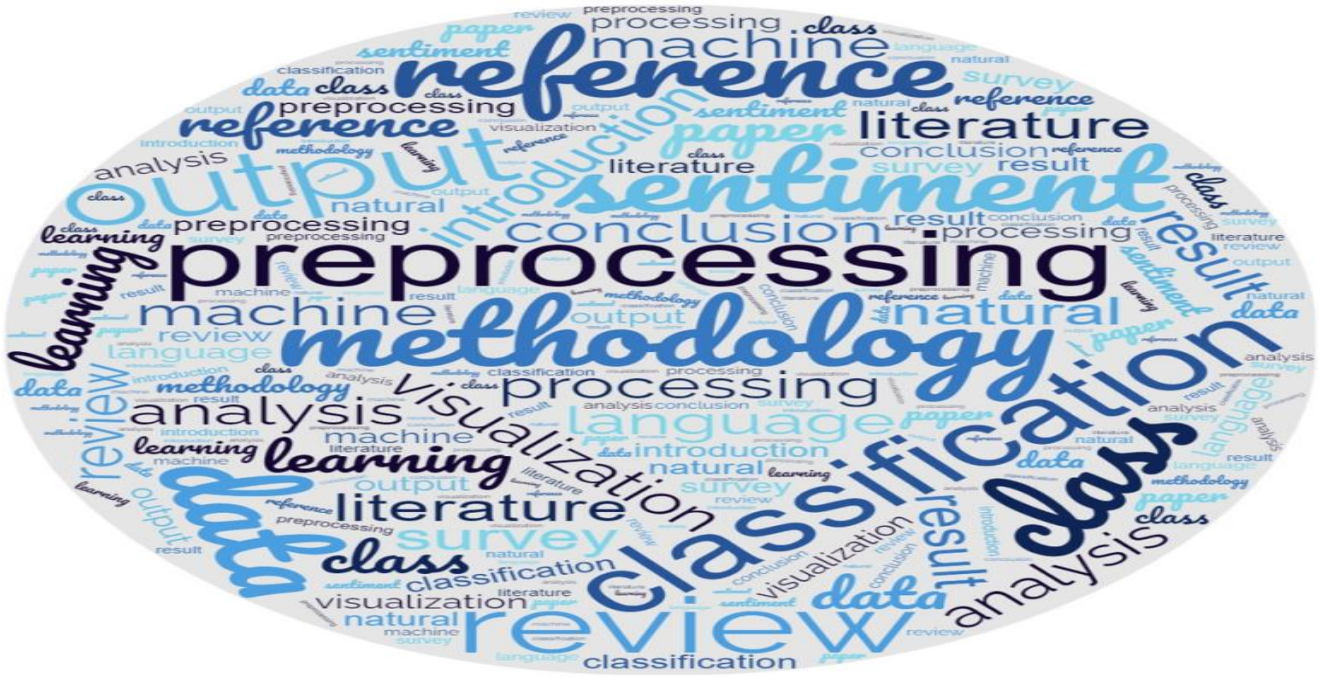


Fig. Word Cloud

4.3 SENTIMENT ANALYSIS

Sentiment Analysis or opinion mining is one of the important tasks of NLP (Natural Language Processing) that has acquired much attention in recent years. The sentiment is a feeling, expression, thought or judgement and using sentiment analysis one can study the target audience's sentiments towards a particular product. It is a form of text analysis that indicates polarity (e.g. a positive or negative opinion) within whole text, sentence or paragraph. Knowing people's emotions is important for companies and first time buyers because consumers can communicate their thoughts and feelings more freely. With the technological improvements in the field of machine learning and automation, companies can create systems that automatically analyzes customers feedback, survey responses and social media interactions. In this way, companies can listen to their customers closely and customize goods and services to suit the needs of their customers

Fig. Categories of Sentiment

4.4 MACHINE LEARNING

ML is being used without specific programming to give computers, the ability to learn. It involves statistical and predictive analysis for allowing the machine to recognize several patterns and to use this knowledge to uncover secret insights into information supplied. Both algorithms for machine learning are split into the supervised machine learning & un-supervised machine learning algorithms. This is accomplished by observing the latest trend and studying how to apply it in a new pattern. On the other hand, unsupervised machine learning algorithms are mainly used to group various specific data types and are applied in various areas where data separation is necessary. In sentiment analysis, classification plays an important role. A pre-classified sample of database which is named a training dataset is used for training and creating a classifier in the classifications stage of sentiments analysis by using machine learning techniques. Upon learning the pattern, this classifier then can label the previous unlabeled data. The precision of a classifier, however, often depends heavily on the data used for training. Therefore, we see that

for sentiment analysis, supervised machine learning methods are better suited. The key machine learning classifier used for sentiments analysis are the following:

4.4.1 Naive Bayesian Classifiers- It is believed that the Naive Bayesian Classifier is very simple and easy in terms of implementation. This is not any single algorithm but consists of the set based on Bayes theorem of various classification algorithms. A term used to define an event's probability. This probabilistic classifier utilizes and analyses all the characteristics present in the vector of the function differently, i.e., it considers them independently of each other. By analyzing a pre-categorized collection of documents, we can learn the pattern.

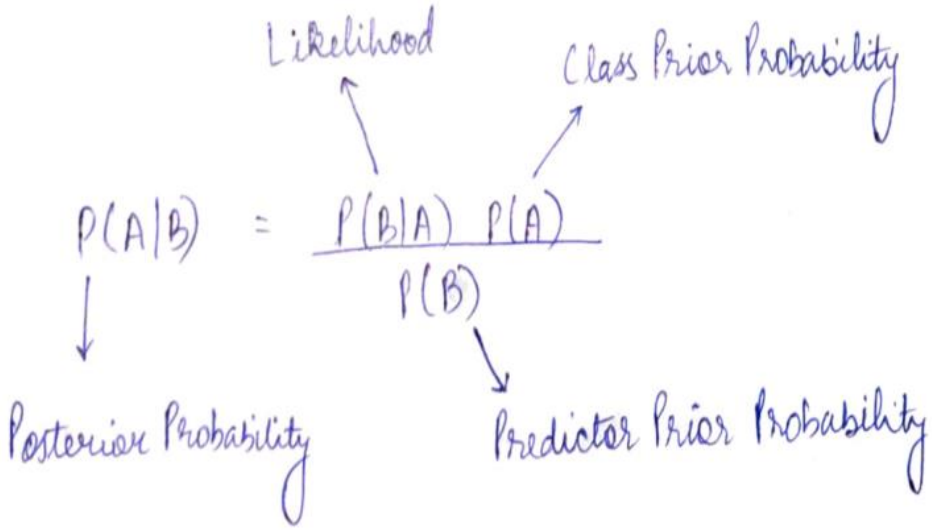


Fig. Formula for Naïve Bayes Classification

This model states that the conditional probabilities of the event P(A) occurring could be determined in presence of the two events, P(A) & P(B) if P(B) has already occurred.. The Naive Bayes classifier's input for training consists of preprocessed data along with its extracted features. The classifications process is conducted on the data set of test data after completing the training and then, depending on the outcomes, the new data. A polarity of the feelings of the data is given by this classification method. For instance, the "It was good" review statement would have resulted in positive polarity.

4.4.2 SVM Classifier-

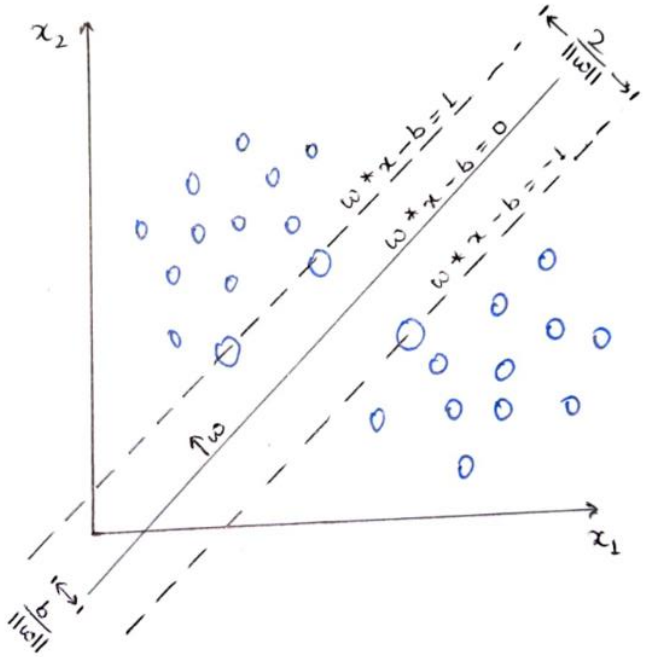


Fig. Graph for SVM

SVM is a popular machine learning technique that employs a statistical approach. It is extremely effective at text classification. There is an n-dimensional space in the SVM, in which n represents number or quantity of features presented in a vector of the function. In the n-dimensional space, each of the data elements presents in the training dataset is registered, the value of each character is the coordinate value.

In this particular n-dimensional space, the key concept of this approach is to find linear separators that best differentiate the various groups. SVM uses a differentiation function with the following parameters: "X" is the vector of the function; the weight vector is "w", and the bias vector is "b". On the training set, the weights & preload vector are automatically learned. Between these two classes, a margin that is far from a document is described. The classifier margins are defined by this distance and indecisive choices are reduced by maximizing this margin. While some features are important to this system, due to the sparse nature of the text, they are correlated and therefore well suited for SVM text classifications.

4.4.3 Decision Tree- For classification issues, the decision trees are mainly used. Depending on the important trait or attributes, also known as the independent variable, the tree is split. Based on these attributes, the space of training data is described in hierarchical form. There is a condition for each attribute value, which is the presence or the absence of one or more than one words. The inner nodes are labeled with characteristics, however, the edges that exit the nodes are called a trace of the weight of the dataset. The name of every leaf in the tree was a group or class.. In this way, in inferring what value is required of the element, a decision tree classifier associates data from an element.

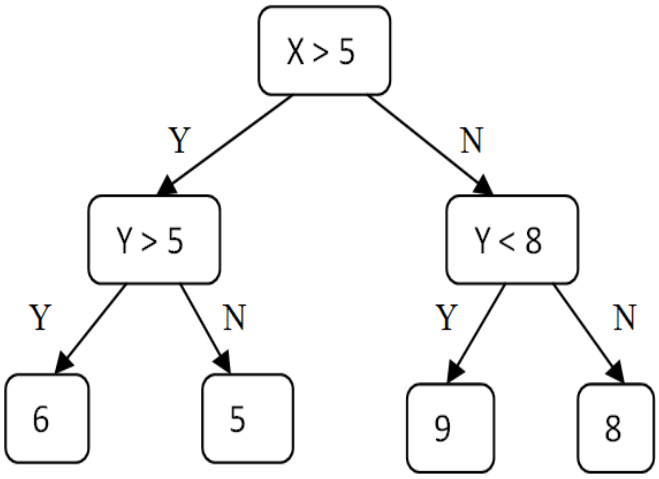


Fig. Decision Tree

4.5 PRODUCT RECOMMENDATION

A product recommendation system is a solution that provides relevant product suggestions to the customers in real time. It is a powerful data filtering platform that depends on algorithms, artificial intelligence, machine learning, and other data analyzing practices. It is a concatenation collecting, storing, analyzing, and filtering customers' data to provide highly personalized relevant products to each and every customer. Relevant products meet the customers' requirements, tastes, and preferences. The quality of data should be very high to achieve such refined targeting at an individual level. But most importantly, we need the right tool to understand the customer data and business needs.

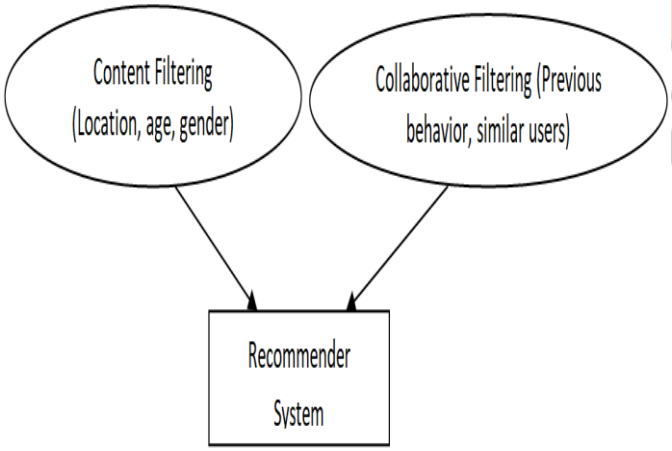


Fig. Model for Recommendation of the Product

V. CONCLUSION

Sentiment analysis or opinion mining is a field of study that analyzes people's sentiments, attitude or emotions towards certain entities. This paper tackles with a fundamental problem of sentiment analysis, sentiment polarity categorization and hence, provides recommendation for the product. Users can see a visual representation of reviews instead of scrolling down and reading all the reviews of a product one by one. Hence, users can save their time and effort for finding what they want to purchase. A particular product is recommended if the polarity of the content of the review is positive or neutral, otherwise, the product will not be recommended to the user.

REFERENCES

[1] Raj Sinha, "Data analysis and sentiment analysis on Amazon reviews", International Journal for Research in Applied Science and Engineering Technology [IJRASET], volume-9, pp. 2200-2206, 2021.
[2] Arwa S. M. AlQahtani, "Product sentiment analysis for Amazon reviews", International Journal of Computer Science and Information Technology [IJCSIT], volume 13, pp.15-30, 2021.
[3] Somsurva Dutta and Santosh Bothe, "Analysis of Amazon reviews using machine learning approach", International Journal for Research in Applied Science and Engineering Technology [IJRASET], volume-9, pp. 313-323, 2021.
[4] Xing Fang and Justin Zhan, "Sentiment analysis using product review data", Journal of Big Data [JBD], 2015.
[5] Waqar Muhammad, Khurum Nazir Junejo, Maria Mushtaq and Muhammad Yaseen Khan, "Sentiment analysis of product reviews in the absence of labeled data using supervised learning approaches", Research Gate, 2019.
[6] Najma Sultana, Sourabh Chandra, Pintu Kumar and Sk Safikul Alam, "Sentiment analysis for product review", Research Gate, 2019.
[7] Pravesh Kumar Singh, "Analytical study of feature extraction techniques in opinion mining", Research Gate, pp.85-94, 2013.
[8] Minu P Abraham and Udaya Kumar Reddy, "Feature based sentiment analysis of mobile product reviews using machine learning techniques", International Journal of Advanced Trends in Computer Science and Engineering [IJATCSE], volume-9, pp. 2289-2296, 2020.
[9] Tanjim Ul Haque, Nudrat Nawal Saber and Faisal Muhammad Shah, "Sentiment analysis on large scale Amazon product reviews", IEEE-International Conference of Innovative Research and Development [ICIRD], 2018.
[10] Raheesa Safrin, K.R.Sharmila and T.S.Shri subangi, "Sentiment analysis on online product review", International Research Journal of Engineering and Technology [IRJET], volume -4, pp. 2381-2388, 2017.
[11] Rajkumar S Jagdale, Vishal S Shrisat and Sachin N Deshmukh, "Sentiment analysis on product reviews using machine learning techniques", Springer, pp 639-647, 2018.
[12] T.K Shivaprasad and Jyothi Shetty, "Sentiment analysis of product reviews", IEEE, 2017.
[13] Prashant Pandey, Muskan and Nitasha Soni, "Sentiment analysis on customer feedback data", IEEE, 2019.
[14] Monir Yahya Ali Salmony and Arman Rasool Faridi, "Supervised sentiment analysis on Amazon product reviews", IEEE, 2021.
[15] Anjana Madhav C and Lavanya M, "Sentiment analysis of product reviews for overall product rating", IEEE, 2020.
[16] Duvvuru Mahammad Dawood Khan, "Sentiment analysis of product based reviews", [IJIRT], volume-8, pp. 467-473, 2021.
[17] Panthati Jagadeesh, Ranga Tarun Kumar, Challa Manish Reddy and Jasmine T. Bhaskar, "Sentiment analysis of product reviews", Research Gate, 2018.
[18] P Rakesh, M Sandeep and G Jagadeesh, "Amazon product review sentiment analysis using machine learning", International Research Journal of Computer Science [IRJCS], volume-8, pp.136-141, 2021.
[19] Arpita Lasod and Rahul Pawar, "Sentiment analysis using machine learning techniques", International Journal of Innovative Research in Technology [IJIRT], volume-6, pp.153-157, 2019.
[20] K Ashok Kumar, "Sentiment analysis of Amazon product reviews using machine learning", Research Gate, volume-82, pp.5245-5254, 2020.
[21] Kiran Shehzadi and Usman Ahmed Raza, "Sentiment analysis by using deep learning and machine learning techniques", International journal of Advanced Trends in Computer Science and Engineering [IJATCSE], volume-10, pp.754-761, 2021.
[22] Sobia Wassan, Xi Chen, Tian Chen, Muhammad Waqr and N Z Jhanjhi, "Amazon product sentiment analysis using machine learning techniques", Research Gate, volume 30, pp.695-703, 2021.
[23] Vineet Jain and Mayuri Kambli, "Amazon product reviews: Sentiment analysis", Research gate, 2020.
[24] Jyoti Budhwar,"Sentiment analysis based method for Amazon product reviews", International Journal of Engineering Research and Technology [IJERT], volume-9, pp.54-57, 2021.
[25] Sayyed Johar and Samara Mubeen, "Sentiment analysis on large scale Amazon product reviews", International Journal of Science Research in Computer Science and Engineering [IJSRCSE], volume-8, pp.07-15, 2020.
[26] P Rakesh, M Sandeep and G Jagadesh, "Amazon product review sentiment analysis using machine learning", International Research Journal of Computer Science [IRJCS], volume-08, pp.136-141, 2021.

EVALUATION AND VISUAL REPRESENTAION OF ONLINE PRODUCTS USING MACHINE LEARNING BASED SENTIMENTAL ANALYSIS

Surabhi Agarwal

P G Student Department of CSE Integral University, Lucknow, Uttar Pradesh, India.
surabhim@student.iul.ac.in

.**Mohd Usman Khan**

Assistant Professor Department of CSE Integral University, Lucknow, Uttar Pradesh, India.
usmankhan@iul.ac.in

ABSTRACT

Owing to the rise in demand for e-commerce with people preferring online buying of goods and products, there is huge amount information being shared. The e-commerce websites are carrying very big volume of data. Also, social media helps a great hand in sharing of this information. This has greatly influenced consumer preferences all over the world. Due to the intense reviews provided by the customers, there is a feedback environment being developed for helping customers buy the right product and guiding companies to enhance the features of product suiting consumer's demand. The only disadvantage of availability of this large volume of data is its range and its structural non-uniformity. The customer finds it difficult to precisely find the review for a particular feature of a product that user intends to buy. Also, there is a mixture of positive and negative reviews thereby making it difficult for customer to find a satisfactory response. Also these reviews suffer from fake reviews from fake users. So to avoid this confusion and make this review system more transparent and user friendly we propose a technique to extract feature based opinion from a diverse hub of reviews and processing it further to differentiate it with respect to the aspects of the product and further categorize it into positive and negative reviews using machine learning based approach. Decision making on both individual and organizational level is always accompanied by the search of other's opinions on the same because data holds expressed opinions and sentiments. The volume, variety and velocity are the key properties of this data. There are several tools and algorithms available to perform sentiment detection and analysis, which are better than unconventional, time consuming and error prone methods used earlier.

KEYWORDS

Sentiment Analysis, Opinion Mining, Text Analysis, Product Review, Polarity Detection, Machine Learning, Feature Extraction.

1. INTRODUCTION

In the recent years E-Commerce has increased suddenly in size everywhere in the world, and most of the population is preferring to buy products through these websites. As a result, large amount of data in the form of reviews is produced which helps expecting buyers to choose the right product. Furthermore these reviews contain opinionated contents which can be useful for the company to identify the areas which need to be work on. However it is not possible for the user to read each and every review about the product. Moreover, reading only few reviews may not present the exact idea about the product. It is quite possible that some of the reviews lack information sources, which the users have no way to differentiate. Besides the reviews and ratings provided a little information to get the specific features of the product. Due to all the above problems, the user is unable to make a fully informed decision about the product. Opinion mining is also known as sentiment analysis that can be used to extract customer reviews from different sources on the internet. This technique uses various algorithms to analyze the large volume of data and gather sense out of it. This technique helps to identify the orientation of a sentence thereby recognizing the positive and negative elements in it. Automated opinion mining can be implemented by using a machine learning based approach. Opinion mining uses natural language processing to extract the subjective information from the data entered by the customers.

The enormous amount of competition to attract and maintain customers online is fascinating businesses to implement novel strategies to enhance the customer experiences. It is becoming compulsory for companies to examine customer reviews on online platforms such as Amazon to understand better how customers rate their products and services. The purpose of this study is to investigate how companies can conduct sentiment analysis based on Amazon reviews to gain more intuitions into customer experiences. The dataset selected for this research consists of customer reviews of Amazon products, which

enables a business person to gain insights on customer reviews regarding specific product and services. The study will enable companies to pinpoint the reasons for positive and negative reviews, followed by effective strategies to address them accordingly. The aim of this research is to help companies to use sentiment analysis to understand customer experiences and customers to understand whether a particular product is to be purchased or not.

2. METHODOLOGY

2.1 PROPOSED METHODOLOGY

The dataset used for this project is from the amazon.com. The reviews in the dataset are consists of the attributes such as: Reviewer ID, Product ID, Review Text, Rating and time of the review. The main source of data used is the product reviews from Amazon. The reviews of a **JBL Digital Sound Bar** have been obtained by building a web crawler. The web crawler has been written in Python using a scrapping library. Along with the review text, some additional data related to the reviews such as reviewer name, review date, overall rating and comments were also obtained. The crawler is called periodically to get the most up-to-date reviews. Each review is generally treated as a sentence or a group of sentences. They are cleaned and stored in a .CSV file on google sheet. The first stage of analysis involves preprocessing of the reviews. Preprocessing involves the following operations: stemming, stop word removal and part-of-speech tagging.

Then, sentiment analysis is performed on the preprocessed reviews and overall sentiment score for each review is generated. Further for feature extraction, there are two cases:

Single Feature – If the review contains only a single feature, then the sentiment score of the review is assigned to the feature.

Multiple Features – Some reviews have multiple features contained in them. So the above procedure will not work in this case. Rules are defined to extract multiple features and assign the correct sentiment score to those features. For reviews containing more than one sentence, first check if the review contains a word from an adjective word set or not. If it does not contain one, then it is assumed to be objective. If it does contain an adjective, the feature that corresponds to that adjective is found by looking for a set of predetermined nouns near that adjective.

2.2 OBJECTIVE

By using this work -

- A customer can get receive recommendation for the product
- A business organization can get the recommendation whether to keep the product or to discard it in future.

2.3 PROBLEM STATEMENT

An application that collects reviews from the users about a certain product and analyzes them. It would segregate the reviews into positive and negative reviews. The negative reviews will be helpful to the companies to further enhance their product based on the user's feedback. The application will provide the pros and cons of the individual feature of the product and hence, reports about the sentiment analysis performed on the products. Then the further aim is to create a recommendation system that recommends products to users according to the feature requirement of user.

There is a strong correlation between user reviews on amazon.com, it is very costly to obtain sentiment labels for large training data and expressions as data of Amazon is unstructured, informal and fast evolving. Amazon has huge data that is present in both structured and unstructured formats. A dataset will be taken and then classified according to its sentiment. It is to collect valuable reviews of a product.

The process of separating emotions, comments from the reviews will begin with feature extraction. The process of labeling begins where the words present in reviews are classified as per five categories, that is **Recommended, Risky, Not Recommended.**

2.4 MODEL ARCHITECTURE

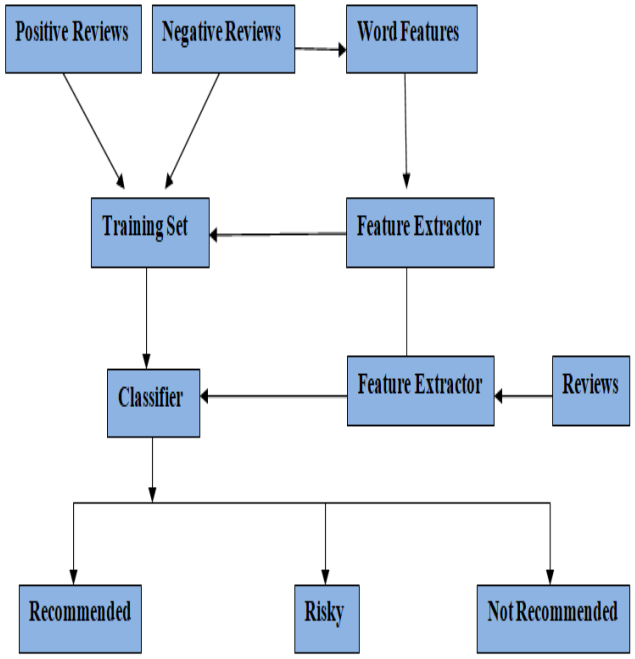


Fig.1- Sentiment Analysis Procedure Flowchart

3. PROPOSED WORK

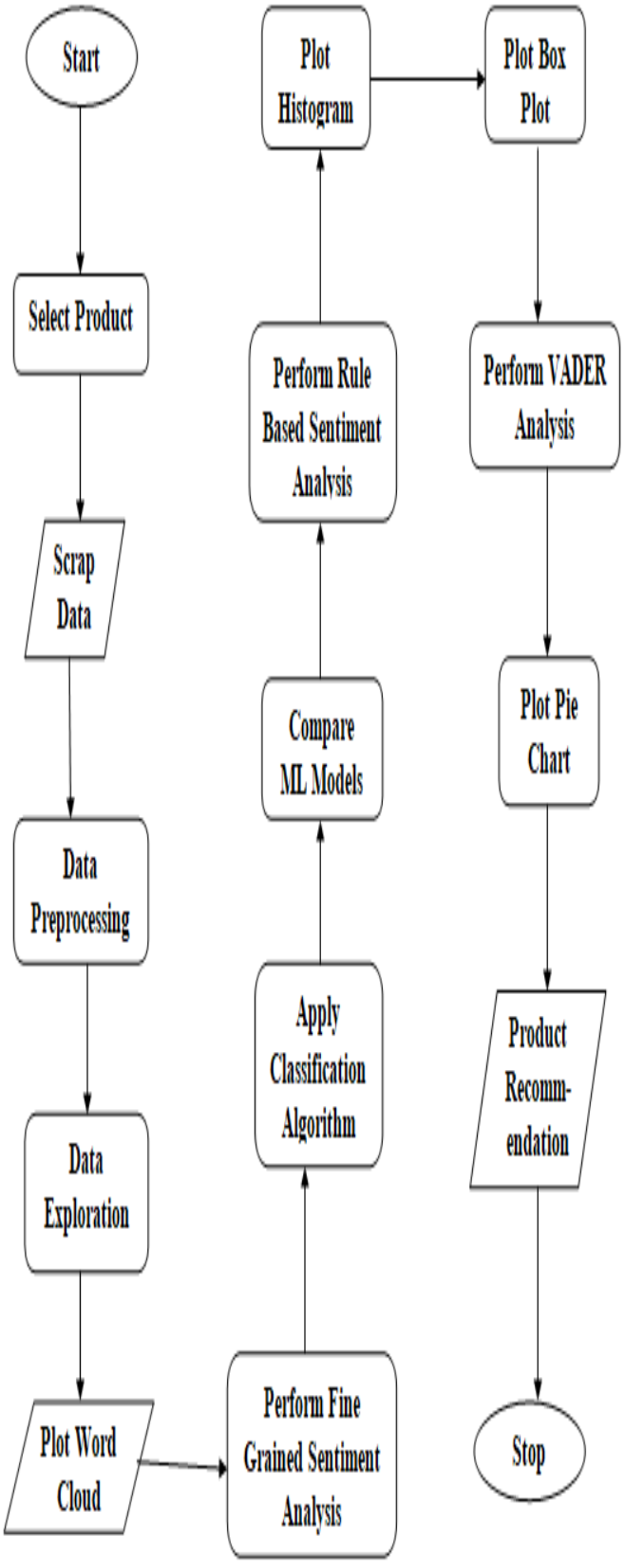
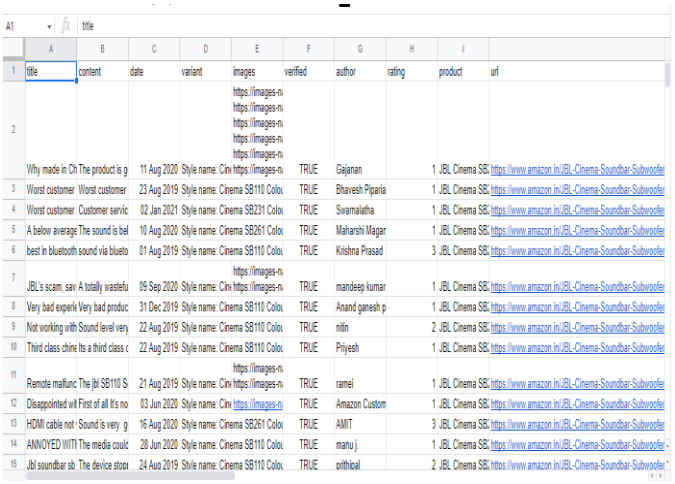


Fig.2- Approach Flowchart

3.1 DATA COLLECTION

The dataset used for this project is taken from the amazon.com. . The reviews of a **JBL Digital Sound Bar** have been obtained by writing a series of code. The code has been written in Python using a scrapping library. The reviews in the dataset are consists of features such as: reviewer ID, product ID, review text, rating and time of the review.



	title	content	date	variant	images	verified	author	rating	product	url
2					https://images-n https://images-n https://images-n https://images-n https://images-n					
	Why made in Ch	The product is g	11 Aug 2020	Style name: Cin	https://images-n	TRUE	Gajanan	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
3	Worst customer	Worst customer	23 Aug 2019	Style name: Cinema SB110 Colou		TRUE	Bhavesh Piparia	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
4	Worst customer	Customer servic	02 Jan 2021	Style name: Cinema SB231 Colou		TRUE	Swarnalatha	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
5	A below average	The sound is bel	10 Aug 2020	Style name: Cinema SB261 Colou		TRUE	Maharshi Magar	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
6	best in bluetooth	sound via blueto	01 Aug 2019	Style name: Cinema SB110 Colou		TRUE	Krishna Prasad	3	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
7	JBL's scam, sav	A totally wastefu	09 Sep 2020	Style name: Cin	https://images-n	TRUE	mandeep kumar	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
8	Very bad experie	Very bad produc	31 Dec 2019	Style name: Cinema SB110 Colou		TRUE	Anand ganesh p	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
9	Not working with	Sound level very	22 Aug 2019	Style name: Cinema SB110 Colou		TRUE	nitin	2	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
10	Third class chine	Its a third class c	22 Aug 2019	Style name: Cinema SB110 Colou		TRUE	Priyesh	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
11	Remote malfunc	The jbl SB110 S	21 Aug 2019	Style name: Cin	https://images-n	TRUE	ramei	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
12	Disappointed wit	First of all It's no	03 Jun 2020	Style name: Cin	https://images-n	TRUE	Amazon Custom	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
13	HDMI cable not i	Sound is very g	16 Aug 2020	Style name: Cinema SB261 Colou		TRUE	AMIT	3	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
14	ANNOYED WITH	The media could	26 Jun 2020	Style name: Cinema SB110 Colou		TRUE	manu j.	1	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer
15	Jbl soundbar sb	The device stopp	24 Aug 2019	Style name: Cinema SB110 Colou		TRUE	prithbal	2	JBL Cinema SB	https://www.amazon.in/JBL-Cinema-Soundbar-Subwoofer

Fig.3- Raw Data

The reviews in the dataset are consists of features such as: reviewer ID, product ID, review text, rating and time of the review.

	title	content	date	variant	images	verified	author	rating	product	url
0	Why made in China ?	The product is good, build quality is premium,...	11 Aug 2020	Style name: Cinema SB261 Colour: Black	https://images-na.ssl-images-amazon.com/images...	True	Gajanan	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...	https://www.amazon.in/JBL-Cinema-Soundbar-Subw...
1	Worst customer care service by JBL	Worst customer care service. Placed a request ...	23 Aug 2019	Style name: Cinema SB110 Colour: Black	NaN	True	Bhavesh Piparia	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...	https://www.amazon.in/JBL-Cinema-Soundbar-Subw...
2	Worst customer care and worst product	Customer service is pathetic. The technical p...	02 Jan 2021	Style name: Cinema SB231 Colour: Black	NaN	True	Swarnalatha	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...	https://www.amazon.in/JBL-Cinema-Soundbar-Subw...
3	A below average sound.	The sound is below average for a 200 watts sou...	10 Aug 2020	Style name: Cinema SB261 Colour: Black	NaN	True	Maharshi Maganti	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...	https://www.amazon.in/JBL-Cinema-Soundbar-Subw...
4	best in bluetooth	sound via bluetooth is excellent but via hdmi ...	01 Aug 2019	Style name: Cinema SB110 Colour: Black	NaN	True	Krishna Prasad	3.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...	https://www.amazon.in/JBL-Cinema-Soundbar-Subw...

Fig.4- Imported Data

Along with the review data, some additional data related to the reviews such as reviewer name, review date, overall rating of the product and comments were also scrapped. The code is called periodically to get the most up-to-date data from the reviews. Each review is usually treated as a sentence or a group of sentences. They are cleaned and stored in a .CSV file on Google sheet. The first step of analysis involves preprocessing of the reviews. Preprocessing involves the

following operations: stemming, stop word removal and part-of-speech tagging. Then, sentiment analysis is performed on the preprocessed reviews and overall sentiment score is generated.

	title	content	verified	rating	product
0	Why made in China ?	The product is good, build quality is premium,...	True	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
1	Worst customer care service by JBL	Worst customer care service. Placed a request...	True	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
2	Worst customer care and worst product	Customer service is pathetic. The technical p...	True	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
3	A below average sound	The sound is below average for a 200 watts sou...	True	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
4	best in bluetooth	sound via bluetooth is excellent but via hdmi ...	True	3.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...

Fig.5- Required Dataset

3.1.2 Tokenization- Token is defined as the minimal unit that a machine understands and processes at a time. All the text strings are processed only after they have passed through tokenization, which is simply the process of splitting the strings into meaningful tokens

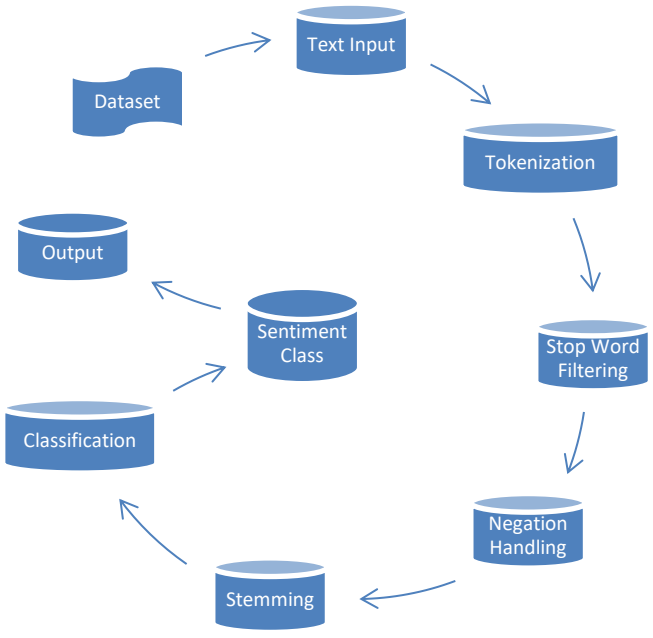


Fig.6- Natural Language Processing Procedure

.

3.1.3 Stop Word Filtering- Stop words are the most commonly occurring words, that often add influence and meaning to the sentences. They act as connectors and their job is to make sure that sentences are grammatically correct. It is one of the conventionally used pre-processing steps across various Natural Language Processing applications. Thus, removing the words that occur commonly in the data is the definition of stop-word removal.

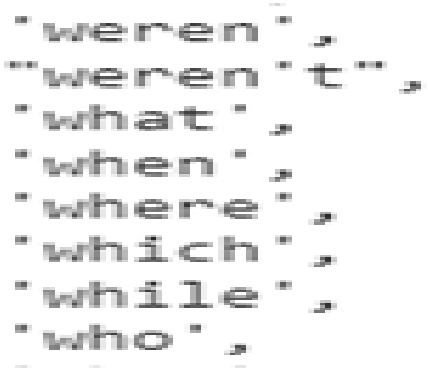


Fig.7- Stop Words

3.1.4 Stemming- Stemming is the process of obtaining the root word from the word given. By using efficient and derived rules, all tokens can be reduced to obtain the root word, also known as the stem. Stemming is a entirely rule-based process through which we put together variations of the token. For example, the word sit will have variations like sitting, sits, sat, etc. It does not make sense to differentiate between sit and sat in many applications, therefore, we use stemming to put both grammatical variances to the root of the word. Stemming is in use for its simplicity. But in the case of Dravidian languages with many more alphabets, and hence, many more permutations and combinations of words possible, the possibility of the stemmer identifying all the variances is very low. In such cases instead of using stemming, we use the lemmatization..

3.1.5 Lemmatization- Lemmatization is a efficient way of converting all the grammatical/inflected forms of the root of the word. Lemmatization makes use of the context and POS tag to determine the inflected form (shortened version) of the word and various normalization rules are applied for each POS tag to get the root word (lemma).

3.1.6 Classification-The most difficult aspect of a system for sentiment analysis is a classifier. Classification is gained in negatives, positive, or neutrals categories. A part of the database is usually used as training sets to generate the classifiers. To a large extent, the accuracy of the classifier depends on the training collection. By using machine learning classifiers like SVM, Bayesian Classifiers and Logistic Regression, the classification can be done. However, before training and testing the classifier, machine learning classifiers perform feature extraction, which can also use deep neural networks for classifying the data.

3.1.7 Result- After the data has gone through the classifier, the output is shown. It shows the polarity of feelings of the whole data, and the degree of detail depends upon the type of

classifiers which is used. The output can be represented in the form of a WORDCLOUD.

Fig.8- Word Cloud

4.3 SENTIMENT ANALYSIS

Sentiment Analysis or opinion mining is one of the important tasks of Natural Language Processing that has gathered much attention in following recent years. The sentiment is a feeling, expression, thought or judgement and using sentiment analysis we can study the target audience's sentiments towards a particular product.

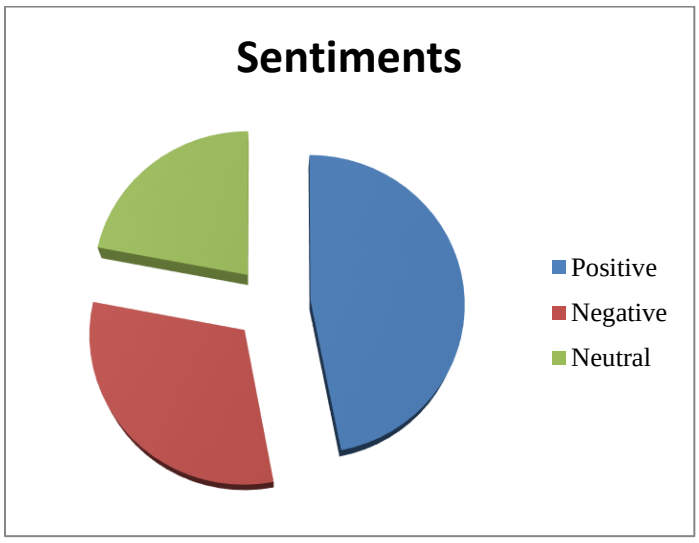


Fig.9- Categories of Sentiment

It is a form of text analysis that points out the polarity (e.g. a positive, negative or neutral opinion) within whole text, sentence or paragraph. Knowing people's emotions is important for companies and first time buyers because consumers can communicate their thoughts and feelings more freely.

With the technological improvements in the field of machine learning and automation, companies can create systems that automatically analyzes customers feedback, survey responses and social media interactions. In this way, companies can listen to their customers closely and customize goods and services to suit the needs of their customers.

4.4 MACHINE LEARNING

Machine Learning is being used without specific programming to give computers, the ability to learn. It involves statistical and predictive analysis for allowing the machine to recognize several patterns and to use this knowledge to uncover secret insights of the information supplied. Both algorithms for machine learning are divided into the supervised machine learning & un-supervised machine learning algorithms. This is obtained by observing the latest trend and studying how to apply it in a new pattern. On the other hand, unsupervised machine learning algorithms are mainly used to group various specific data types and are applied in various areas where data separation is necessary. In sentiment analysis, classification plays an important role. A pre-classified sample of database which is named as a training dataset is used for training and creating a classifier in the classifications stage of sentiments analysis by using machine learning techniques. After learning the pattern, this classifier then can label the previous unlabeled data. The precision of a classifier, however, often depends heavily on the data used for training. Therefore, we see that for sentiment analysis, supervised machine learning methods are better suited. The key machine learning classifier used for sentiments analysis are the following:

4.4.1 Naive Bayesian Classifiers- It is believed that the Naive Bayesian Classifier is very simple and easy in terms of implementation. This is not any single algorithm but consists of the set based on Bayes theorem of various classification algorithms. A term used to define an event's probability. This probabilistic classifier utilizes and analyses all the characteristics present in the vector of the function differently, i.e., it considers them independently of each other. By analyzing a pre-categorized collection of documents, we can learn the pattern.

This model states that the conditional probabilities of the event P(A) occurring could be determined in presence of the two events, P(A) & P(B) if P(B) has already occurred.. The Naive Bayes classifier's input for training consists of preprocessed data along with its extracted features. The classifications process is conducted on the data set of test data after completing the training and then, depending on the outcomes, the new data. A polarity of the feelings of the data is given by

this classification method. For instance, the "It was good" review statement would have resulted in positive polarity.

4.4.2 SVM Classifier- SVM is a popular machine learning technique that employs a statistical approach. It is extremely effective at text classificationThere is an n-dimensional space in the SVM, in which n represents number or quantity of features presented in a vector of the function. In the n-dimensional space, each of the data elements presents in the training dataset is registered, the value of each character is the coordinate value. In this particular n-dimensional space, the key concept of this approach is to find linear separators that best differentiate the various groups. SVM uses a differentiation function with the following parameters: "X" is the vector of the function; the weight vector is "w", and the bias vector is "b". On the training set, the weights & preload vector are automatically learned. Between these two classes, a margin that is far from a document is described. The classifier margins are defined by this distance and indecisive choices are reduced by maximizing this margin. While some features are important to this system, due to the sparse nature of the text, they are correlated and therefore well suited for SVM text classifications.

4.4.3 Logistic Regression- A logistic regression algorithm model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. For example, a logistic regression could be used to predict whether a political candidate will win or lose in an election or whether a high school student will be admitted or not to a particular college. These binary outputs provide direct decisions between two alternatives. Logistic regression has become an important tool in the field of machine learning. It allows algorithms used in machine learning applications to classify input data based on historical data. As additional relevant data comes in as an input, the algorithms get better at predicting classifications within data sets. A logistic regression algorithm model can take into consideration multiple input criteria.

4.5 PRODUCT RECOMMENDATION

A product recommendation system is a solution that provides relevant product suggestions to the customers in real time. It is a powerful data filtering platform that depends on machine learning algorithms, artificial intelligence and other data analyzing practices. It is an addition of collecting, storing, analyzing, and filtering customers' data to provide highly personalized relevant products to each and every customer. Relevant products meet the customers' requirements, needs, tastes, and preferences. The quality of data should be very high to achieve such polished outcome targeting at an individual level. But most importantly, we need right tools to understand the data provided by the customers and business needs.

S NO.	POLARITY	RECOMMENDATION
1	Positive	Recommended
2	Neutral	Risky
3	Negative	Not Recommended

Table 1- Recommendation of the Product

5. RESULT AND ANALYSIS

5.1 DATASET

	title	content	verified	rating	product
0	whi made in china	the product is good build qualiti is premium t...	1	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
1	worst custom care servic by jbl	worst custom care servic place a request for a...	1	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
2	worst custom care and worst product	custom servic is pathet the technic person is ...	1	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
3	a below averag sound	the sound is below averag for a watt sound bar...	1	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
4	best in bluetooth	sound via bluetooth is excel but via hdmi arc ...	1	3.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
...	...	...	...	...	...
718	horribl jbl and amazon	receiv a defaulti product sound issu difficult...	1	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
719	wast of timewast of money	im return two timeswhen i wa purcha but produc...	1	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
720	it better to buy local speaker	speaker not work properli left speaker work an...	1	3.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
721	do not buy wast of money a bit disappoint with...	it been month now and im use thi soundbar with...	1	1.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...
722	not worth it	sound veri muffl	1	2.0	JBL Cinema SB231, 2.1 Channel Dolby Digital So...

723 rows × 5 columns

5.2 WORDCLOUD GENERATED

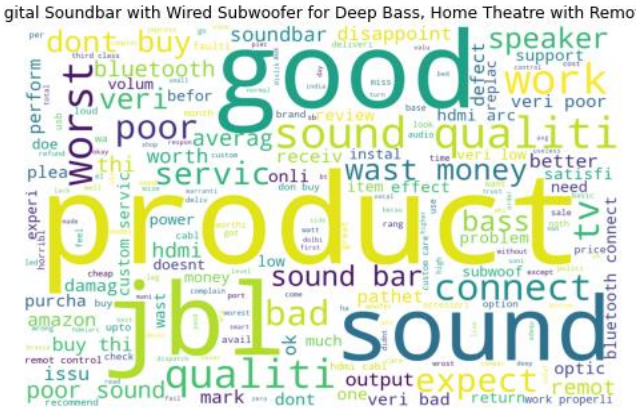


Fig.10- Generated Word Cloud

4.3 POLARITY OF REVIEW

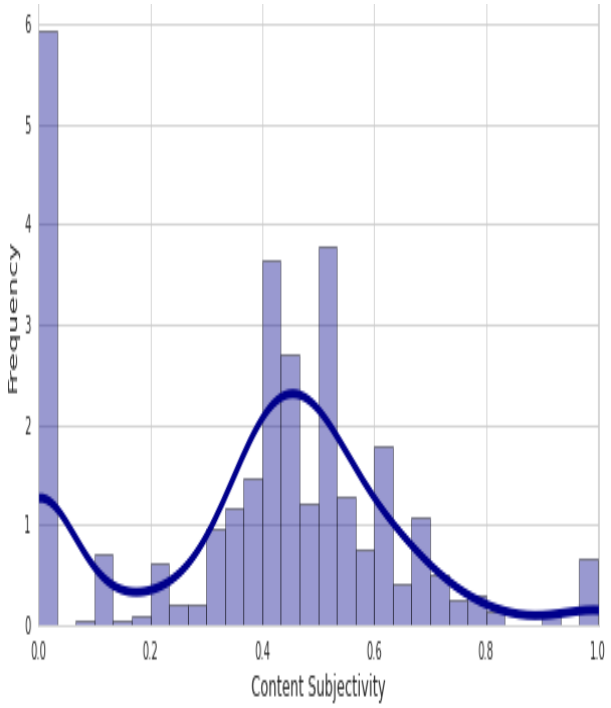


Fig.11- Distribution of Content Subjectivity Score

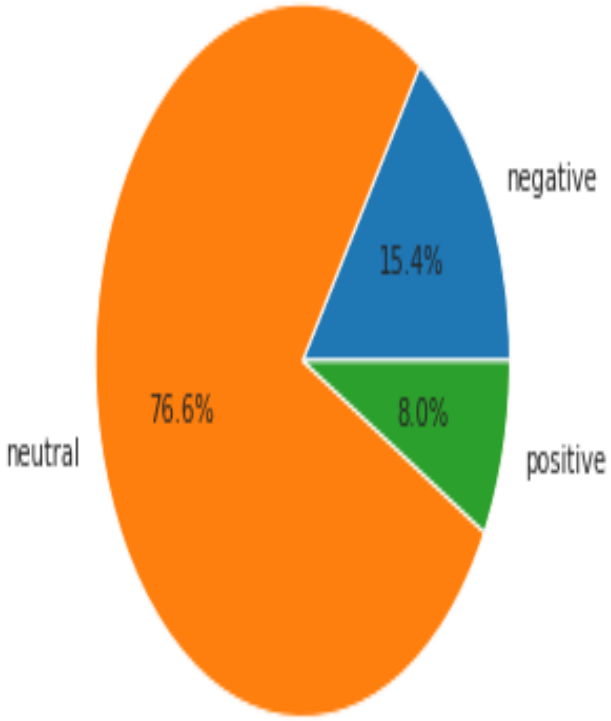


Fig.12- Pie Chart

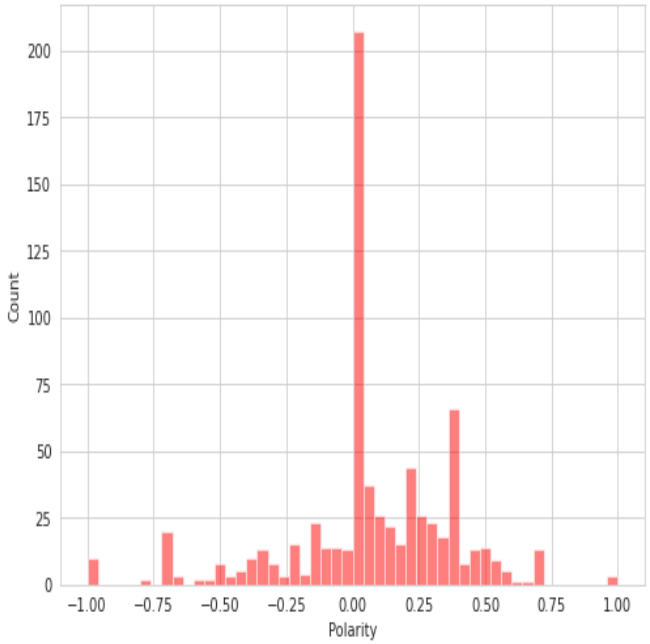


Fig.13- Histogram

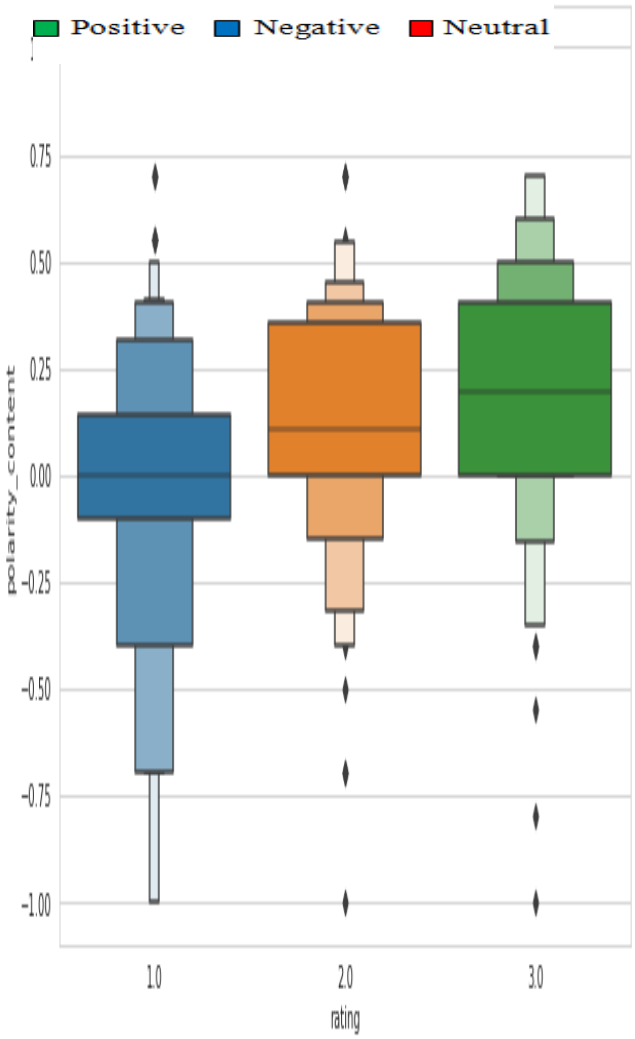


Fig.14- Box Plot

Accuracy Evaluation of ml models

S No.	ML Model	Train Accuracy	Test Accuracy
1	Naïve Bayes	89.43%	89.91%
2	SVM	95.56%	94.91%
3	Logistic Regression	87.13%	88.79%

Table 2- Accuracy Evaluation

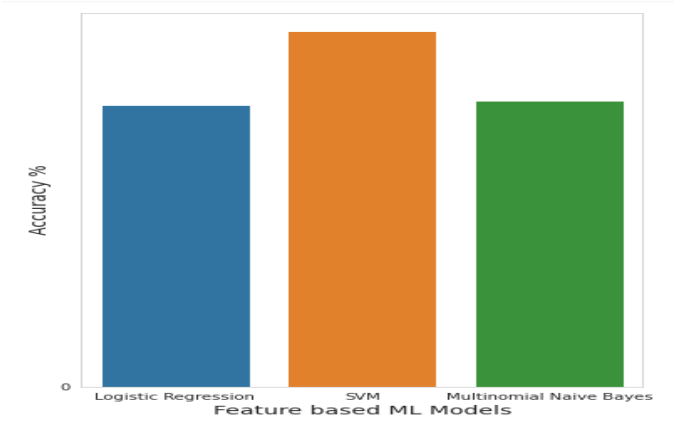


Fig.15- Accuracy of Machine Learning Models

4.4 RECOMMENDATION

According to the result generated, the majority of the user had shown the neutral behavior towards the product, therefore, it is recommended that the product is **risky** to buy.

6. CONCLUSION

With many applications, sentiment analysis is a rapidly growing field. Not only can consumer expectations be fulfilled based on the outcome of the sentiment analysis, but suppliers, distributors, etc. can also get an idea of the user or client's reaction and therefore ensure that they can make and meet the required adjustments. Sentiment analysis has been an important tool for brands looking to learn more about how their customers are thinking and feeling. We have studied different methods and approaches of ML. The techniques of machine learning are much simple and easy to incorporate. These approaches achieve critical outcomes.

The system proposed in this report aims to help a user select a JBL Sound Bar based on his/her needs using review data from previous users. It pulls review data from the Amazon website periodically, processes it and assigns a score to each feature of each phone based on the review data. When the user inputs his/her preferences, the scores are used to determine the best match for the user. This match is guaranteed to be up-to-date.

A model was tested by using Support Vector Machine, Naïve Bayes and Logistic Regression on datasets of product reviews to find the polarity of sentiments and texts whether positive, negative or neutral. The performance resulting models tested to obtain the value of Accuracy, Recall, Precision, and F-1 measure of all three models used. Finally the Support Vector Machine (SVM) algorithm has been achieved higher accuracy, i.e., 90.99% and it is found that the SVM is a robust and better one.

In summary, proposed work tried Naive Bayes, SVM and Logistic Regression. Research is supposed to provide more flexible and accurate solution. The proposed research is supposed to resolve the issue of previous research that was faced during sentiment analysis.

7. FUTURE SCOPE

Amazon is the world's largest e-commerce platform and hence, it is a pool of reactions, sentiments of customers as well. Thus, we can observe their feelings, sentiments, emotions towards the product from this site. Like this, user can opt for any product from this site to get a hand on it's reviews. User just need to copy the URL of the product through the website. So, not only the product used in this work, we can retrieve sentiments of any product just by using product's URL .

Sentiment analysis is a uniquely powerful tool for businesses that are looking to measure attitudes, feelings and emotions regarding their brand. To date, the majority of sentiment analysis projects have been conducted almost exclusively by companies and brands through the use of social media data, survey responses and other hubs of user-generated content. By investigating and analyzing customer sentiments, these brands are able to get an inside look at consumer behaviors and, ultimately, better serve their audiences with the products, services and experiences they offer.

The future of sentiment analysis is going to continue to dig deeper, far past the surface of the number of likes, comments and shares, and aim to reach, and truly understand, the significance of social media interactions and what they tell us about the consumers behind the screens. This forecast also predicts broader applications for sentiment analysis – brands will continue to leverage this tool, but so will individuals in the public eye, governments, nonprofits, education centers and many other organizations.

- Algorithm-Based Sentiment Analysis Plateaus
- Not Just For Marketers and Brands
- Greater Personalization for Audiences
- Deeper, Broader Insights from Sentiment Analysis

REFERENCES

[1] Raj Sinha, “Data analysis and sentiment analysis on Amazon reviews”, International Journal for Research in Applied Science and Engineering Technology [IJRASET], volume-9, pp. 2200-2206, 2021.

[2] Arwa S. M. AlQahtani, “Product sentiment analysis for Amazon reviews”, International Journal of Computer Science and Information Technology [IJCSIT], volume 13, pp.15-31, 2021.

[3] Somsurva Dutta and Santosh Bothe, “Analysis of Amazon reviews using machine learning approach”, International Journal for Research in Applied Science and Engineering Technology [IJRASET], volume-9, pp. 313-323, 2021.

[4] Xing Fang and Justin Zhan, “Sentiment analysis using product review data”, Journal of Big Data [JBD], 2015.

[5] Waqar Muhammad, Khurum Nazir Junejo, Maria Mushtaq and Muhammad Yaseen Khan, “Sentiment analysis of product reviews in the absence of labeled data using supervised learning approaches”, Research Gate, 2019.

[6] Najma Sultana, Sourabh Chandra, Pintu Kumar and Sk Safikul Alam, “Sentiment analysis for product review”, Research Gate, 2019.

[7] Pravesh Kumar Singh, “Analytical study of feature extraction techniques in opinion mining”, Research Gate, pp.85-94, 2013.

[8] Minu P Abraham and Udaya Kumar Reddy, “Feature based sentiment analysis of mobile product reviews using machine learning techniques”, International Journal of Advance Trends in Computer Science and Engineering [IJATCSE], volume-9, pp. 2288-2296, 2020..

[9] Tanjim Ul Haque, Nudrat Nawal Saber and Faisal Muhammad Shah, “Sentiment analysis on large scale Amazon product reviews”, IEEE-International Conference of Innovative Research and Development [ICIRD], 2018.

[10] Raheesa Safrin, K R Sharmila and T S Shri Subangi, “Sentiment analysis on online product review”, International Research Journal of Engineering and Technology [IRJET], volume -4, pp. 2381-2388, 2017.

[11] Rajkumar S Jagdale, Vishal S Shrisat and Sachin N Deshmukh, “Sentiment analysis on product reviews using machine learning techniques”, Springer, pp 639-647, 2018.

[12] T K Shivaprasad and Jyothi Shetty, “Sentiment analysis of product reviews”, IEEE, 2017.

[13] Prashant Pandey, Muskan and Nitasha Soni, “Sentiment analysis on customer feedback data”, IEEE, 2019.

[14] Monir Yahya Ali Salmony and Arman Rasool Faridi, “Supervised sentiment analysis on Amazon product reviews”, IEEE, 2021.

[15] Anjana Madhav C and Lavanya M, “Sentiment analysis of product reviews for overall product rating”, IEEE, 2020.

[16] Duvvuru Mahammad Dawood Khan, “Sentiment analysis of product based reviews”, [IJIRT], volume-8, pp. 467-473, 2021.

[17] Panthati Jagadeesh, Ranga Tarun Kumar, Challa Manish Reddy and Jasmine T. Bhaskar, “Sentiment analysis of product reviews”, Research Gate, 2018.

[18] P Rakesh, M Sandeep and G Jagadeesh, “Amazon product review sentiment analysis using machine learning”, International Research Journal of Computer Science [IRJCS], volume-8, pp.136-141, 2021.

[19] Arpita Lasod and Rahul Pawar, “Sentiment analysis using machine learning techniques”, International Journal of Innovative Research in Technology [IJIRT], volume-6, pp.153-157, 2019.

[20] K Ashok Kumar, “Sentiment analysis of Amazon product reviews using machine learning”, Research Gate, volume-82, pp.5245-5254, 2020.

[21] Kiran Shehzadi and Usman Ahmed Raza, ”Sentiment analysis by using deep learning and machine learning techniques”, International journal of Advanced Trends in Computer Science and Engineering [IJATCSE], volume-10, pp.754-761, 2021.

[22] Sobia Wassan, Xi Chen, Tian Chen, Muhammad Waqr and N Z Jhanjhi, ”Amazon product sentiment analysis using machine learning techniques”, Research Gate, volume 30, pp.695-703, 2021.

[23] Vineet Jain and Mayuri Kambli, ”Amazon product reviews: Sentiment analysis”, Research gate, 2020.

[24] Jyoti Budhwar,“Sentiment analysis based method for Amazon product reviews”, International Journal of Engineering Research and Technology [IJERT], volume-9, pp.54-57, 2021.

Surabhi Agarwal

by Prakriti Mishra

Submission date: 20-Jul-2022 01:13AM (UTC-0500)

Submission ID: 1866839648

File name: Final_Thesis_Surabhi.docx (4.02M)

Word count: 10469

Character count: 56289

Surabhi

ORIGINALITY REPORT

6%	4%	2%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	docplayer.net Internet Source	4%
2	"Intelligent Information and Database Systems", PubliCation	

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		